

# Predicting Hydrophobic Solvation by Molecular Simulation: 1. Testing United-atom Alkane Models

Miguel Jorge<sup>1\*</sup>, Nuno M. Garrido<sup>2</sup>, Carlos J. V. Simões<sup>3,4</sup>, Cândida G. Silva<sup>3,5</sup>,  
Rui M. M. Brito<sup>3,5</sup>

*<sup>1</sup>Department of Chemical and Process Engineering, University of Strathclyde, 75 Montrose Street, Glasgow G1 1XJ, United Kingdom*

*Email – miguel.jorge@strath.ac.uk*

*<sup>2</sup>LSRE – Laboratory of Separation and Reaction Engineering – Associate Laboratory LSRE/LCM, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal*

*<sup>3</sup>Chemistry Department and Coimbra Chemistry Centre, Faculty of Science and Technology, University of Coimbra, 3004-535 Coimbra, Portugal*

*<sup>4</sup>BSIM<sup>2</sup> – Drug Discovery, Parque Tecnológico de Cantanhede, 3060-197 Cantanhede, Portugal*

*<sup>5</sup>Center for Neuroscience and Cell Biology, University of Coimbra, 3004-504 Coimbra, Portugal*

**Abstract:** We present a systematic test of the performance of three popular united-atom force fields – OPLS-UA, GROMOS and TraPPE – at predicting hydrophobic solvation, more precisely at describing the solvation of alkanes in alkanes. Gibbs free energies of solvation were calculated for 52 solute/solvent pairs from Molecular Dynamics simulations and thermodynamic integration making use of the IBERCIVIS volunteer computing platform. Our results show that all force fields yield good predictions when both solute and solvent are small linear or branched alkanes (up to pentane). However, as the size of the alkanes increases, all models tend to increasingly deviate from experimental data in a systematic fashion. Furthermore, our results confirm that specific interaction parameters for cyclic alkanes in the united-atom representation are required in order to account for the additional excluded volume within the ring. Overall, the TraPPE model performs best for all alkanes, but systematically underpredicts the magnitude of solvation free energies by about 6% (RMSD of 1.2 kJ/mol). Conversely, both GROMOS and OPLS-UA systematically overpredict solvation free energies (by ~13% and 15%, respectively). The systematic trends suggest that all models can be improved by a slight adjustment of their Lennard-Jones parameters.

## 1 - Introduction

The solvation free energy is defined as the free energy difference required to bring a solute molecule from the gas phase to the solution phase at infinite dilution [1]. It determines the equilibrium solubility of gases in liquids and plays an important role in accurately predicting the equilibrium solubility of solid phases [2]. It is also a crucial component in determining how a given solute partitions between two different liquid solvents, through the partition coefficient [3], and how different ligands bind to protein active sites [4]. Thus, the prediction of solvation free energies assumes great importance in many fields, including pharmaceuticals, oil and gas, environmental, and bioengineering [5]. Such predictions are usually carried out using empirical, group contribution or statistical correlation approaches, but molecular simulation methods have appeared as attractive alternatives due to their potential for increased accuracy and transferability [6]. However, their effectiveness relies on the existence of accurate and predictive molecular models. Because water is ubiquitous in biological systems and widely used in industrial and environmental processes, it is not surprising that the overwhelming majority of studies of solvation free energy have focused on water as a solvent – i.e., they have attempted to describe the hydration free energy [7-9]. However, many important processes involve the interaction of solutes with both hydrophobic and hydrophilic environments (e.g., protein-ligand binding, solvent extraction), or even with mostly hydrophobic media (e.g., in the oil and gas industry). As such, increasing attention needs to be paid to the ability of molecular models to predict the hydrophobic component of solvation free energies.

In the general case, the solvation free energy can be broken down into three components: i) the free energy required to build a cavity of the size and shape of the solute within a bulk solvent, normally termed the *cavity formation cost*, mainly dictated by repulsive forces and solvent reorganization entropy; ii) the *dispersion contribution*, arising from favorable interactions of the London type between solute and solvent; iii) the *electrostatic contribution*, arising from (favorable or unfavorable) polar interactions between solute and solvent molecules. This decomposition of the solvation free energy is by no means universal, but it is convenient from the theoretical point of view, as a correspondence can be established between the different components and the parameters of the molecular model (for example, between the dispersion contribution and the Lennard-Jones parameters). In fact, most calculation schemes for solvation free energies using molecular simulation separately compute the electrostatic term before calculating the dispersion and cavity terms together [7-9]. Unfortunately, there is no experimental counterpart to the individual components, and the quality of a molecular model has to ultimately be judged against the overall solvation free energy. Furthermore, these separate components are not thermodynamic functions of state, so the contribution of each term will depend on the path used to calculate it. Thus, for most systems of interest, it is not possible to decouple the effect of the electrostatic parameters (point charges, multipole expansions, etc.) from the dispersion-repulsion parameters (e.g.,

Lennard-Jones). This fact, coupled with the large number of degrees of freedom of the solute and solvent models, introduces significant complexity in the parameterization of molecular models for solvation free energy predictions. In this paper, we circumvent these problems by focusing on a particular subset of practically relevant solute-solvent systems where electrostatic interactions are negligible – alkanes dissolved in alkanes.

There are comparatively few molecular simulation studies that report solvation free energy predictions in non-aqueous solvents, and even fewer that deal with completely non-polar solvents. In fact, a very recent comprehensive benchmark study of molecular models for solvation free energy predictions did not include any alkane as a solvent [10]. A relevant exception is the development of recent versions of the GROMOS force field, which made use of solvation free energies in cyclohexane, as well as in water, as part of their validation [11] and parameter fitting [12] efforts. They focused on amino acid side-chain analogs as solutes, given the availability of experimental data for comparison [13]. Overall, they concluded that version 43a2 of the force field led to solvation free energies that were too high [11], which prompted a later reparameterisation and extension of the model [12]. Interestingly, in this later reparameterisation the authors proposed two different versions of the force field, one optimized for pure liquid properties (version 53A5) and another for solvation free energy calculations (version 53A6) [12].

Among the amino acid side-chain analogs, four turn out to be completely non-polar alkane molecules – methane, propane, butane and isobutane. Solvation of these solutes in cyclohexane is thus governed purely by hydrophobic (dispersion) interactions, and thus it is interesting to consider this subset of solutes in more detail. Villa and Mark [11] reported solvation free energy predictions for those alkanes in cyclohexane that were in reasonable agreement with experiment, albeit systematically too high (i.e., solvation was less favorable). A recent study by Szklarczyk et al. [14] reports excess free energies, which are related to the self-solvation free energies, for a few alkane molecules using GROMOS 45A3 parameters, observing good agreement with experimental data. MacCallum and Tieleman [15], and later Chang et al. [16], carried out similar comparisons using the OPLS-AA force field [17] and also obtained solvation free energies of alkanes in cyclohexane that were systematically too positive compared to experiment. A recent study by our team of solvation free energies of solutes and solvents with different polarity found that the solvation free energy of propane in n-hexadecane was predicted very well by the TraPPE force field [18]. These few studies highlight the strong dependence of predictive performance on the quality of the molecular model, even for simple molecules such as alkanes. This warrants a more thorough and extensive assessment of the ability of current models for predicting hydrophobic solvation.

In this paper, we rigorously test three popular united-atom (UA) force fields – OPLS-UA [19], GROMOS [20] and TraPPE [21, 22] – against an extensive database of experimental Gibbs energies of solvation [23, 24] for their ability to predict the hydrophobic component of the solvation free energy, without the added complexity of worrying about the

electrostatic component. This allows us to shed new light on the solvation mechanisms in non-polar solvents, which can be used to inform the development of more general models for polar molecules. In a companion publication [25], we propose an improved UA model that provides highly accurate solvation free energy predictions for these systems. Using molecular dynamics (MD) and thermodynamic integration (TI), we have computed the solvation free energy of over 150 solute-solvent combinations including linear, branched and cyclic alkanes. Because these calculations are computationally intensive, we made use of the resources available in an Iberian volunteer computing platform called IBERCIVIS [26]. This platform operates in similar ways to other popular citizen distributed computing platforms such as SETI@home [27] or Folding@home [28]. Each volunteer participating in the project donates their personal computer's idle times, during which it receives a work package from the central IBERCIVIS server and returns the calculation output after completion [29, 30].

## 2 – Computational Methods

### 2.1 Solvation free energy calculations

Solvation free energies were calculated by the thermodynamic integration (TI) method [31] based on a series of molecular dynamics (MD) simulations carried out using the GROMACS software, version 4.5.4 [32]. Our approach has been described in detail in previous publications [33-36]; here we will briefly describe the general features of the method and focus on particular implementation issues. In the current implementation of TI in GROMACS, the Hamiltonian of the solution (in our case, containing a single solute molecule in a three-dimensional periodic box of solvent) is decomposed into solute-solute, solute-solvent and solvent-solvent contributions. A coupling parameter,  $\lambda$ , is applied to the solute-solvent part of the Hamiltonian, so that its contribution can be modulated between full interactions (corresponding to  $\lambda=0$ ) and no interactions ( $\lambda=1$ ). This particular implementation obviates the need for a separate calculation of a single solute in vacuum to account for the contribution of intramolecular solute-solute interactions, via a thermodynamic cycle [36]. Essentially, the solute is made to gradually “disappear” from the solution using the coupling parameter. This approach assumes that the conformational space explored by the solute molecules is the same in vacuum and in the solution. To confirm that this is indeed the case, we have calculated probability distributions for the dihedral angles in long hydrocarbon chains in the gas phase and in the solvated state. Figure S1 shows the results for the longest molecule simulated here, hexadecane, in which we can see that the distributions are quite similar in the two phases. The percentage of trans configurations is slightly higher for the more compressed liquid state (75%) than for the gas phase (71%), but the difference is not significant. For shorter alkanes, the differences will be even smaller.

In TI, independent simulations are carried out for different values of  $\lambda$ , between 0 and 1, and the gradient of the Hamiltonian with respect to  $\lambda$  is averaged over a large number of

equilibrated configurations. The solvation free energy ( $\Delta G_{\text{sol}}$ ) is then calculated by numerically integrating the Hamiltonian gradient over  $\lambda$ , following equation (1):

$$\Delta G_{\text{sol}} = \int_0^1 \left\langle \frac{\partial H(\mathbf{p}, \mathbf{q}, \lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda \quad (1)$$

where  $H$  is the Hamiltonian of the system, which depends on the particle positions ( $\mathbf{q}$ ) and momenta ( $\mathbf{p}$ ), as well as on the coupling parameter.

It was demonstrated previously [37, 38] that the accuracy of the free energy is affected by numerical integration errors when the standard trapezoidal rule is applied. Here, we fit the data for the Hamiltonian gradient as a function of  $\lambda$  to a theoretically-based expression, which is then integrated analytically (for details, the reader is referred to ref [37]). We have used a total of 15  $\lambda$  points, which is more than sufficient to minimize integration error [37]. This conservative choice was motivated by our additional interest in analyzing trends in the variation of the Hamiltonian gradient plots, as described below. Note that because our systems involve non-polar alkanes described at the UA level, we only need to consider the Lennard-Jones contribution to the solvation free energy. The standard procedure for charged molecules involves a first TI step where the molecule is “uncharged”, followed by the decoupling of the solute-solvent LJ interactions [34]. For the particular case of UA alkanes, where all sites are electronically neutral, the first step is unnecessary.

Each individual MD simulation was performed in the NpT ensemble, using a Langevin thermostat [39] to control the temperature at 298 K (except where explicitly noted) and a Parinello-Rahman barostat [40] to keep the pressure at 1 bar. The equations of motion were integrated using the leapfrog algorithm [41] with a time step of 2 fs. Dispersion interactions were handled using a switched cutoff between 1.0 and 1.1 nm. Long-range dispersion corrections were added to energy and pressure. Full technical details about the calculations have been provided in our previous papers [33-37]. It is worth noting that special care was taken to ensure that the results did not suffer from finite-size effects, particularly for the larger alkane molecules – cubic box lengths were at least 3.5 nm, and as large as 5 nm for hexadecane.

## 2.2 Implementation of GROMACS on the IBERCIVIS platform

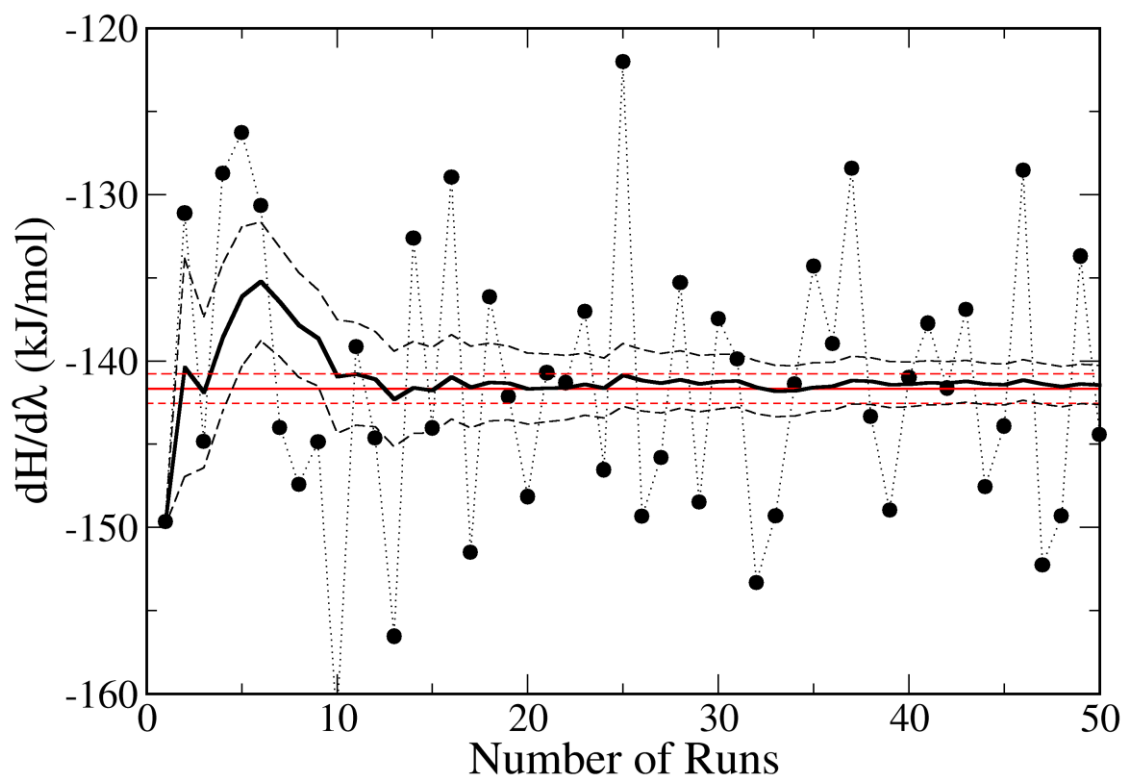
The IBERCIVIS platform is designed around the non-commercial Berkeley Open Infrastructure for Network Computing (BOINC) framework [42], aimed at deploying volunteer and grid computing projects. BOINC operates under the client-server architecture and the workflow is reasonably straightforward. A client program runs on the volunteer’s computer under client-predefined settings. The local computer contacts with any of the project servers it is registered with to retrieve a processing job (*workunit*). Once the processing of the downloaded *workunit* is concluded, the client contacts the project server to upload the result and requests a new job from the same or an alternative server.

The implementation of GROMACS version 4.5.4 on IBERCIVIS involved the use of so-called wrapper programs, which allow applications (or sequences of applications) to be launched in their native environment, while ensuring the use of the appropriate options for creating *workunits* and file templates. In this work, multiple wrappers bundled with applications belonging to the GROMACS machinery were deployed on the master BOINC server. These wrappers are responsible for preparation and analyses of simulations. Since GROMACS (in particular *mdrun*) is a wave-aware program, i.e. allowing checkpoints, a queue system based on *bash* scripts was implemented to handle the generation, continuation, assimilation and validation of *workunits*.

In practice, the usual TI approach involves performing a single long simulation at each value of  $\lambda$  to guarantee equilibration and adequate statistical sampling of the Hamiltonian gradient. In a distributed computing platform like IBERCIVIS, however, jobs are farmed to a large number of separate computational nodes (the volunteers), as described above. In such a platform, it is convenient to split large jobs into smaller portions to minimize disruption and loss of data (we experienced on average a loss of about 3% of jobs submitted for computation). As such, we have opted to carry out a large number of small calculations for each value of  $\lambda$  instead of a single large calculation. After testing, we decided to use 50 simulations of 200 ps each, giving a total of 10 ns per point and 150 ns for each solvation free energy calculation. Each individual simulation was started from a different initial configuration, extracted from a pre-equilibrated solute-solvent trajectory (each configuration was sampled evenly from the last 4 ns a full 5 ns trajectory). We confirmed that 200 ps were sufficient to ensure convergence of the Hamiltonian gradient for each individual configuration, as shown in Figure S2. The first 20 ps of each simulation were discarded and the Hamiltonian gradient was sampled over the remainder of the trajectory.

The second test was to ensure that the Hamiltonian gradient for a given value of  $\lambda$  converged to the expected value after averaging over all 50 individual configurations. In Figure 1, we plot the individual values of the gradient for each simulation (points) together with their running average (solid black line). This is compared against the result of a single long simulation at the same value of  $\lambda$  using the same total computational time (i.e., 10 ns). It is clear that after averaging over about 20 runs, the results are statistically equivalent to those of the long run, thus validating our procedure. This was confirmed for several values of  $\lambda$  and different solute-solvent pairs; the comparison shown in Figure 1 is for a particularly stringent test case. Our overall assessment showed that 25 to 30 configurations were sufficient to ensure statistically converged results. However, we have decided to submit 50 configurations for each pair in order to account for potential losses (as mentioned above, about 3% of the submitted jobs crashed and returned no results) and to increase the precision of the calculated free energies (notice the decrease in the statistical error in Figure 1 as the number of configurations increases). From the results of the 50 individual runs, a standard error for the Hamiltonian gradient at each  $\lambda$  value was determined. The uncertainty estimate in the final

solvation free energy values, reported as error bars in the remainder of this paper, was obtained by error propagation of the individual standard errors over the analytical integration procedure [37].



*Figure 1 – Convergence of the Hamiltonian gradient (in kJ/mol) with respect to the coupling parameter  $\lambda$  as a function of the number of individual configurations used. All data is for pentylbenzene in water at  $\lambda = 0.75$ . Circles connected by thin dotted lines show the individual data points for each MD run starting from a different initial configuration. The full black line shows the running average of those data points, and the dashed black lines on either side show the average  $\pm$  the standard error. This is compared to the result of a single long 10 ns simulation, shown as the horizontal red line ( $\pm$  standard error).*

## 2.2 Molecular models

In this paper, we tested three of the most popular united-atom models for alkanes – OPLS-UA [19], GROMOS [20] and TraPPE [21, 22]. As mentioned above, by adopting a UA representation for the alkane molecules, we effectively treat each molecule as being electronically neutral, which is a good approximation for the vast majority of alkanes. This approach reduces the electrostatic contribution to the free energy to zero, thus simplifying the calculations. More importantly for our purposes, it reduces the number of fitting parameters when developing the model (see the second paper in this series [25]). In this work, we are mainly concerned with the performance of the non-bonded part of the model, i.e., the Lennard-Jones potential (see details in Supplementary Material). Although there are some small differences in bonded terms among the three models, all of them provide an accurate description of the structure of alkane molecules [19, 20]. As such, the effect of those differences in the free energy predictions is of relatively minor importance. Apart from the

actual values of the Lennard-Jones parameters, the three force fields differ in their parameterization strategy, including the choice of atom types used, the order in which these were parameterized and the target experimental data used in the parameterization and subsequent validation. Because these choices can significantly affect the performance and transferability of each model, we describe them briefly here. The reader is referred to the original publications for further details [19-22].

OPLS is the oldest of these models, and was parameterized in 1984 [19] to match the densities and vaporization enthalpies of a series of small alkanes, which included linear, branched and cyclic molecules. The authors began by fitting the Lennard-Jones parameters for the CH<sub>2</sub> site against liquid cyclopentane. They then used the same parameters for CH<sub>2</sub> groups of linear alkanes, and fitted the parameters of CH<sub>3</sub> groups in linear molecules against n-butane. These parameters were able to predict experimental properties of other small linear alkanes, but not of ethane. As a consequence, the authors chose to employ a distinct set of parameters for CH<sub>3</sub> groups belonging to ethane. Parameters for CH and C groups belonging to branched alkanes were fitted to properties of liquid isobutane and neopentane, respectively. However, once again the authors were unable to find appropriate parameters for those sites without also changing the CH<sub>3</sub> parameters. This led to the use of distinct parameter sets for CH<sub>3</sub> groups belonging to single- and double-branched alkanes. To complete the parameter set, the authors used the values of Verlet and Weis [43] for united-atom methane. The complete set of OPLS-UA alkane parameters is provided in Table 1.

*Table 1 – Lennard-Jones parameters for the three tested United-Atom alkane force fields.*

Site	OPLS-UA		GROMOS		TraPPE	
	$\sigma$ (nm)	$\varepsilon$ (kJ/mol)	$\sigma$ (nm)	$\varepsilon$ (kJ/mol)	$\sigma$ (nm)	$\varepsilon$ (kJ/mol)
CH <sub>4</sub>	0.373	1.2301	0.371	1.26361	0.373	1.2305
CH <sub>3</sub> (linear)	0.3905	0.7322	0.3748	0.86715	0.375	0.81482
CH <sub>3</sub> (ethane)	0.3775	0.86609	0.3748	0.86715	0.375	0.81482
CH <sub>3</sub> (single-branch)	0.3910	0.66944	0.3748	0.86715	0.375	0.81482
CH <sub>3</sub> (double-branch)	0.3960	0.60668	0.3748	0.86715	0.375	0.81482
CH <sub>2</sub> (linear)	0.3905	0.49371	0.40704	0.41054	0.395	0.38247
CH <sub>2</sub> (cyclopentane)	0.3905	0.49371	0.3955	0.47928	0.388	0.46808
CH <sub>2</sub> (cyclohexane)	0.3905	0.49371	0.3955	0.47928	0.391	0.43649
CH <sub>2</sub> (cycloheptane)	0.3905	0.49371	0.3955	0.47928	0.390*	0.43025*
CH <sub>2</sub> (larger cycles)	0.3905	0.49371	0.3955	0.47928	0.389	0.42401
CH	0.385	0.33472	0.5019	0.09489	0.468	0.08314
C	0.380	0.2092	0.664	0.007	0.640	0.004157

\* Interpolated linearly between values for cyclohexane and cyclooctane

The argument used by Jorgensen and co-workers to start their parameterization by fitting properties of a cyclic alkane was that these are the only class of alkanes composed of



CH<sub>2</sub> sites alone. This would then avoid the need to simultaneously fit parameters for more than one type of site against a larger set of experimental data points, thus significantly simplifying the fitting procedure. However, it has been shown subsequently [44, 45] (and corroborated by our own results – see below) that in order to achieve a sufficiently realistic representation of both linear and cyclic molecules in a UA representation, distinct parameters should be employed for CH<sub>2</sub> groups belonging to each class of molecules. This is because parameters for cyclic molecules need to account for the additional excluded volume within the ring and for the reduced exposure of ring sites to surrounding molecules [44]. To make matters worse, cyclopentane turns out to be a particularly troublesome molecule, due to its highly strained configuration [46]. We will return to these points below when we compare the performance of the three models in predicting solvation of cyclic alkanes.

The GROMOS force field was first presented in 1984 [47], but the most recent parameter set for UA alkanes is from 2001 [20] (the 45a3 parameter set). It was also fitted against densities and enthalpies of vaporization of pure liquids, but included hydration free energies in the validation step. Contrary to the approach used in OPLS-UA, Schuler et al. [20] began by determining the parameters for CH<sub>2</sub> and CH<sub>3</sub> groups in linear molecules by simultaneously fitting them to match the density and enthalpy of vaporization of alkane chains from ethane to eicosane (chain length of 20 carbon atoms). Interestingly, they reported an inability to match both the density and vaporization enthalpy of ethane, although they decided this was an acceptable compromise. The authors then used a distinct set of parameters for CH<sub>2</sub> groups in cyclic alkanes, leading to a good match for the liquid properties of cyclohexane. Although predictions for smaller cyclic alkanes were of lower quality, the authors assigned greater importance to predicting the behavior of cyclohexane due to its extensive use as a solvent in practical applications. Finally, Schuler et al. [20] simultaneously fitted parameters for CH and C groups against density and vaporization enthalpies of several branched alkanes. The GROMOS 45a3 parameter set for alkanes was completed by adopting methane parameters that were virtually identical to those of OPLS-UA.

The TraPPE model is the most recent of the three models tested here (the first paper dates from 1998 [21]), and used a completely different set of target experimental data for fitting and validation – vapor/liquid equilibrium (VLE) properties. Methane parameters in TraPPE were taken from the OPLS-UA model because they led to a good prediction of the two-phase diagram and critical properties of that molecule. Martin et al. [21] began by fitting the CH<sub>3</sub> parameters against VLE data for ethane. They then transferred those parameters to the other linear alkanes and adjusted the CH<sub>2</sub> parameters to match the VLE properties of n-octane. The model was validated by accurately predicting the vapor/liquid coexistence curve of n-pentane, as well as critical properties of several other linear alkanes. In a subsequent paper, Martin et al. [22] fitted parameters for CH and C groups by matching experimental VLE data of isobutane and neopentane, respectively. In a later paper, Lee et al. [45] attempted to predict the coexistence properties of cyclooctane using the parameters

determined for linear CH<sub>2</sub> groups, but were unable to do so. They therefore developed a new set of parameters for cyclic CH<sub>2</sub> sites. However, the authors noted that the performance of the model deteriorated as the size of the cyclic molecule decreased (namely for cyclohexane and cyclopentane). Quite recently, Keasler et al. [46] derived new parameters independently for each of those smaller cyclic alkanes, leading to the full parameter set shown in Table 1.

A particularly important issue when comparing different force fields is the need for consistency in the simulation parameters. Because a particular model is parameterized using a particular set of simulation parameters, its accuracy may deteriorate when those parameters are changed. A pertinent example in this context is the recent reparameterization of point-charge models of water using Ewald summations to handle long-range electrostatic interactions [48, 49]. As Ewald calculations are quite time consuming and were not routinely used at the time, original water models were calibrated using a simple spherical cutoff radius for Coulomb interactions [50]. Later, when Ewald sums were used together with the original parameters, it was shown that the accuracy of the original models deteriorated [48]. In our particular systems, electrostatic interactions play no role, but an analogous point can be made about the treatment of dispersion interactions. In the case of the three models studied here, OPLS-UA and TraPPE both used a spherical cutoff of 1.4 nm together with long-range dispersion corrections applied to energy and pressure [19, 21]. However, GROMOS was parameterized using a spherical cutoff of 1.4 nm and a neighbor list cutoff of 0.8 nm to increase computational expediency (the so-called twin range cutoff) [20]. Crucially, no long-range dispersion corrections were applied. In this paper, we chose to compare all models on an equal basis, so the same technical parameters were used in all simulations (switched cutoff between 1.0 and 1.1 nm plus long-range corrections). It is thus important to establish the effect of that choice of parameters, if any, on the resulting calculations. We carried out such tests for solvation free energies, the main objective of this paper, and for bulk solvent density, for which calculations are much faster.

In Figure S3, we plotted the calculated density of nonane as a function of cutoff radius using the TraPPE model (the qualitative trends are expected to be model-independent). It is clear that when long range corrections are used (black line), the resulting density is practically unaffected by the actual cutoff radius, at least beyond a value of 0.9 nm. We also compared the solvation free energy of hexane in hexadecane obtained using our choice of cutoff scheme and using a larger cutoff of 1.4 nm (always with long-range corrections), and found the results to be within statistical error of each other. This confirms the outcome of a more detailed test of the effect of simulation parameters on the precision of free energy calculations, which concluded that a cutoff radius of 0.9 nm can be safely used (thus reducing computational cost), provided long-range corrections are employed [51]. Furthermore, we confirmed that the choice of cutoff scheme (simple spherical cutoff, switched cutoff or twin range cutoff) also had no statistically relevant impact on the density or solvation free energies of alkanes when long-range corrections were applied. Conversely, when long-range

corrections were not used (red line in Figure S3), the resulting bulk density showed a strong dependency on the value of the cutoff radius. As the radius increased, the density tends to converge to the value obtained using long-range corrections, but this limit was not reached even for a cutoff radius as large as 1.8 nm. This shows that models parameterized with long-range dispersion corrections are much more robust in the sense that they can be safely used with alternative set-ups. However, models like GROMOS, which were parameterized without long-range corrections, should be used with care if the cutoff scheme is changed. We will return to this point later when comparing the performance of the different models.

### 3 - Results and discussion

#### 3.1 – Linear Alkanes

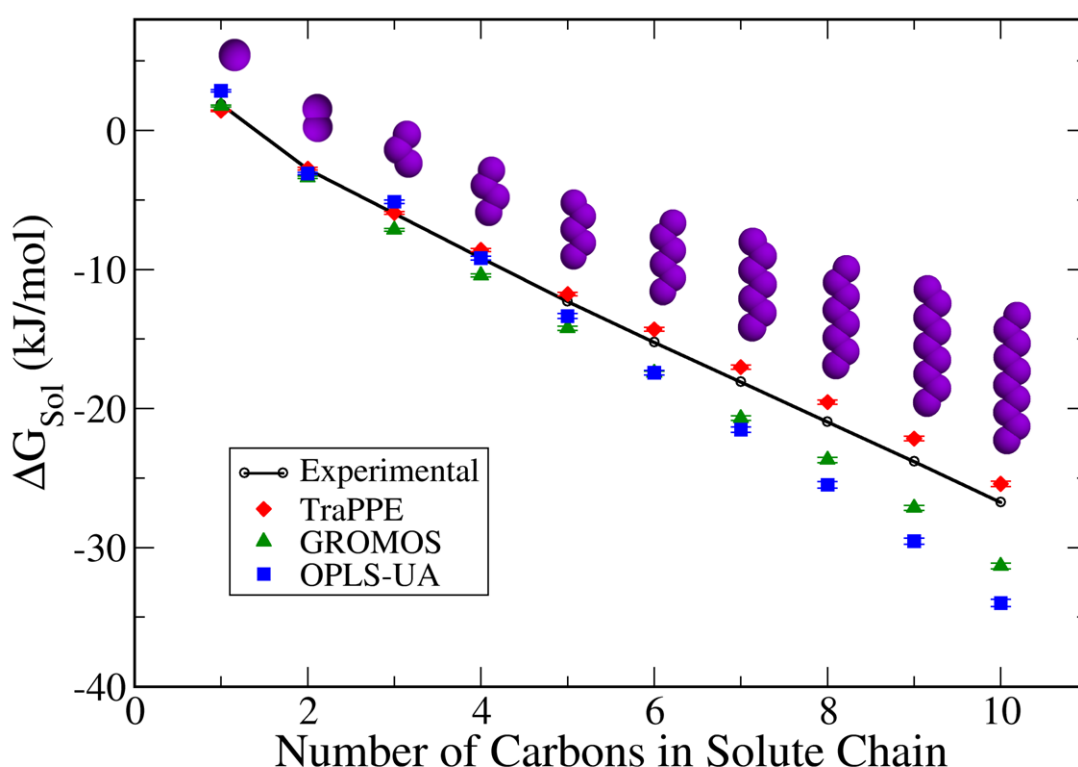


Figure 2 – Solvation free energy of linear alkanes of increasing chain length in hexadecane solvent. The open circles represent the experimental data [23, 24] while the filled points represent simulations using different models – TraPPE, red diamonds; GROMOS, green triangles; OPLS-UA, blue squares. The black line through the experimental points is a guide to the eye. The purple molecules are illustrations of the solute models.

We begin by comparing the performance of the three selected UA models to predict the solvation of linear alkanes of increasing chain length in hexadecane solvent. Hexadecane is a widely used solvent in chemical processes, and is often used as a model compound for the aliphatic components of crude oil [52]. It is also the solvent for which more data points are available in the experimental database [24]. Figure 2 plots the predictions of the three models against experimental data. All models perform quite well for small linear solutes, up

to pentane, but increasingly deviate from experiment at high solute chain lengths. In the case of both GROMOS and OPLS-UA, this deviation is in the direction of overestimated solvation – i.e., the alkanes are somewhat too soluble in hexadecane, and their free energies of solvation are systematically too negative. TraPPE also shows a systematic, albeit smaller, deviation, but in this case it underestimates the magnitude of the free energy. This suggests that the Lennard-Jones parameters of GROMOS and OPLS-UA are probably too attractive to deal with large alkane molecules, while those of TraPPE are slightly underestimating the attractive dispersion term.

Analyzing the data in a bit more detail shows that the predictions for methane and ethane using the three models are statistically indistinguishable. This is somewhat expected, as the parameters for these molecules are practically identical in all three models. For methane, this was somewhat intentional, as the parameters can all be traced back to the same original source [43]. However, the parameters for CH<sub>3</sub> groups in ethane were all calibrated independently, sometimes using distinct experimental properties. It is somewhat reassuring that very similar parameters were obtained in all three cases (see row for “CH<sub>3</sub> (ethane)” in Table 1) and that they lead to accurate solvation free energy predictions, indicating the robustness of the ethane UA parameters. However, the same cannot be said for the parameters in larger linear molecules, which show more significant differences between the three models (see Table 1). In particular, the CH<sub>2</sub> (linear) parameters in OPLS-UA are significantly more attractive than for the other two models –  $\sigma$  is smaller and  $\epsilon$  is larger – and it compensates for this fact by employing a less attractive set of parameters for CH<sub>3</sub> groups in larger linear molecules (much larger  $\sigma$  and smaller  $\epsilon$  than in the other models). This is a direct consequence of the choice made in OPLS-UA to parameterize CH<sub>2</sub> groups against cyclic alkanes. This compensation works well for the smaller alkanes (up to pentane), against which the model was originally parameterized. However, as the chain length increases, the importance of CH<sub>2</sub> groups dominates and the model becomes overall too attractive.

The differences between TraPPE and GROMOS parameters are more subtle, but the latter has systematically more attractive parameters for both CH<sub>2</sub> and CH<sub>3</sub> groups in linear alkanes (Table 1). This may at least partially be justified by the choice of cutoff scheme employed during the parameterization of each model. Because GROMOS did not apply long range dispersion corrections, part of the attractive interactions in the bulk liquid are “missing” when compared with TraPPE. To compensate for this, GROMOS makes use of more attractive LJ parameters than TraPPE in order to obtain an accurate description of bulk liquid properties. The drawback of this, as we have shown earlier, is that the model becomes less robust to changes in the simulation parameters. Furthermore, as we will show later, the differences in cutoff scheme are not enough to account for the discrepancy between simulation and experiment.

Another interesting observation in Figure 2 is that the free energy increases linearly with chain length for solutes larger than ethane, both in experimental data and in simulations.

Methane, however, is an exception in that it is less soluble than one would anticipate based on this linear trend. It is also the only alkane to display a positive (i.e. unfavorable) solvation free energy in hexadecane. This exception to the trend becomes even clearer when one examines the raw data from the thermodynamic integration method (Figure 3 for the TraPPE model). Starting from ethane and increasing the solute chain length involves the addition of a single  $\text{CH}_2$  group within the alkane chain. The excluded volume of the additional site, and hence the disruption to the solvent structure, is essentially the same regardless of the size of the chain, and the same can be said about the additional interactions between the  $\text{CH}_2$  site and the solvent. As a consequence, both the cavity formation term and the dispersion contribution change monotonically with solute chain length. Reflecting this fact, all the Hamiltonian gradient plots for ethane and beyond cross over at the same point, around  $\lambda=0.75$  (see Figure 3). On the other hand, the change from methane to ethane is more significant – replacing one hydrogen atom by a terminal  $\text{CH}_3$  group. Thus, the Hamiltonian gradient plot for methane deviates from the observed trend by having a much larger cavity formation cost, in relative terms. So much so, that this unfavorable contribution more than compensates the favorable dispersion interactions, yielding a positive overall solvation free energy.

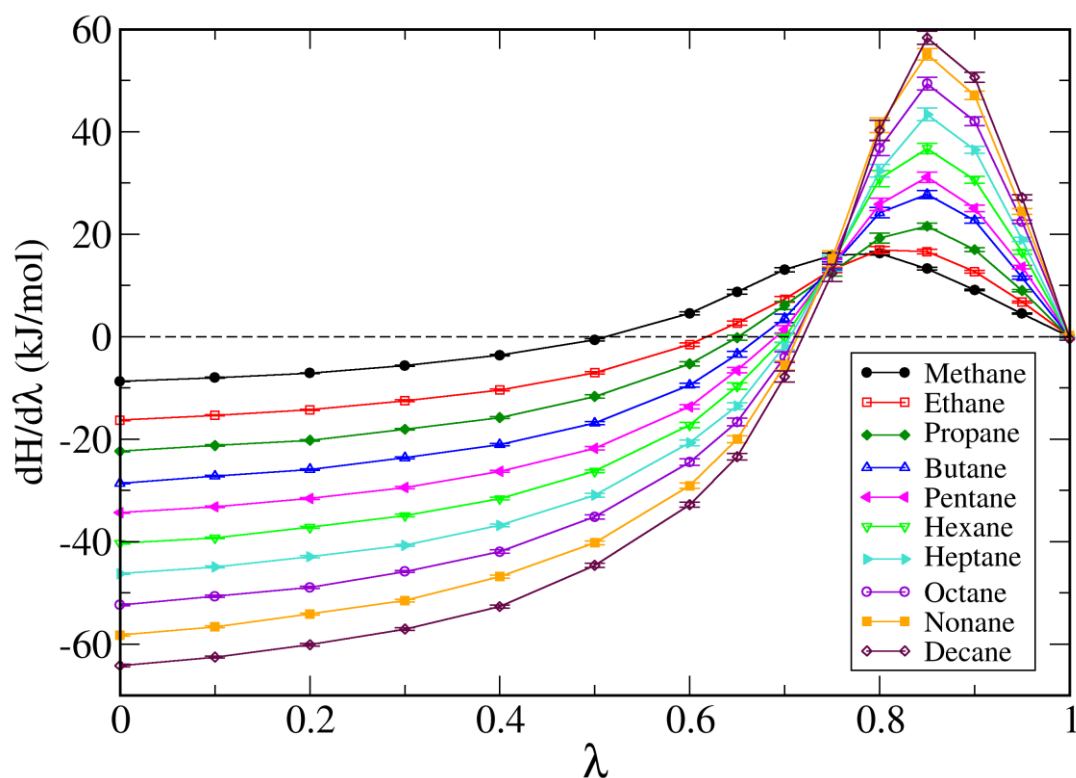


Figure 3 – Gradient of the Hamiltonian with respect to the coupling parameter ( $\lambda$ ) for linear solutes of increasing chain length in hexadecane solvent, obtained from molecular simulations using the TraPPE model.

In Figure 4, we compare the self-solvation free energy – i.e., both solute and solvent are the same alkane – for alkanes of increasing chain length. Once more, the free energy increases linearly with chain length, and all models capture this trend, as expected. Similarly

to what was observed in Figure 2, TraPPE slightly underestimates solvation, while both GROMOS and OPLS-UA systematically overestimate it. The plot of the Hamiltonian gradients for these systems, shown in Figure S4, reveals another interesting trend – all plots cross over at precisely zero. Interestingly, when the plots are normalized by the molecular weight of the solute/solvent, they nearly collapse onto a single master curve (see Figure S5). This suggests that the relative contribution of the cavity and dispersion contributions to the solvation free energy is independent of the chain length, at least for relatively long alkane chains.

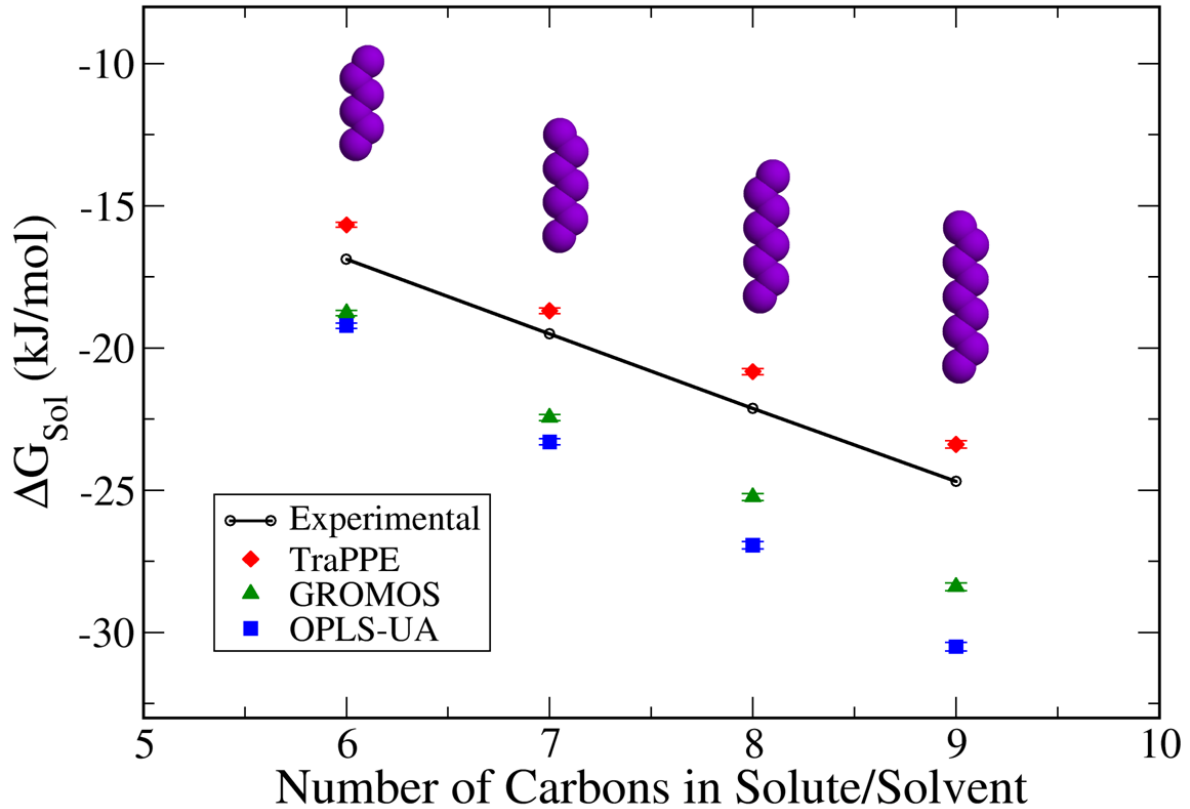


Figure 4 – Solvation free energy of linear alkanes for solute/solvent pairs with increasing chain length (i.e., both the solvent and the solute are the same alkane). Color code is the same as in Figure 2.

It is interesting to interpret these results in light of the direct connection between the pure component saturated vapor pressure ( $P_{sat}$ ) and the self-solvation free energy, through equation (2) [53]:

$$P_v^{Sat} = P^0 M_L \exp\left(\frac{\Delta G_{sol}}{RT}\right) \quad (2)$$

where  $P_v^{Sat}$  is the equilibrium vapor pressure of the fluid,  $P^0$  is the pressure of an ideal gas at a molarity of 1 mol/l,  $M_L$  is the molarity of the pure liquid,  $R$  is the ideal gas constant and  $T$  is the temperature. This equation assumes ideal gas behavior and takes the same standard state in the vapor and liquid phases [53]. The link implied in equation (2) means that any model

that overestimates the vapor pressure of a fluid will also overestimate its self-solvation free energy, and vice-versa. While parameterizing the TraPPE model, Martin et al. [21] concluded that it was not possible to describe all phase equilibrium properties simultaneously using a single set of UA parameters. They then decided to optimize their parameters to describe the critical temperature and orthobaric liquid densities as closely as possible, sacrificing agreement for other properties if necessary. One of the consequences of this choice was that the TraPPE parameters systematically overestimate the saturated vapor pressures of alkanes [21]. Through equation (2), an overestimated vapor pressure leads to self-solvation free energies that are too positive. This agrees precisely with our observations (Figure 4), and suggests that the systematic deviations in the TraPPE predictions are due the decision to sacrifice agreement with experimental vapor pressures.

The final test for solvation free energies of linear alkanes was to keep the solute molecule fixed and change the chain length of the solvent chain. The results of this test for the three models are shown in Figure 5. As observed previously, TraPPE underestimates the magnitude of the free energy, while GROMOS and OPLS-UA overestimate it. However, all three models qualitatively capture the trend of increasing solvation free energy (i.e., less favorable solvation) with increasing solvent chain length observed in the experimental data. In contrast to what was observed when the solute chain length increased, the effect of changing the solvent is relatively minor (notice the difference in scale between Figure 5 and Figure 2). Examining the raw data of the simulations (Figure S6), we can see that the main difference between the curves is in the cavity formation term, which gets slightly larger as the solvent chain length increases. These results can be loosely interpreted as a decreasing probability of formation of rod-like voids (i.e., of the shape of a hexane solute in this particular example) within solvents of progressively larger chain length. In agreement with this interpretation, the solvation free energy levels off, eventually tending towards a value corresponding to solvation in an infinite-chain length solvent.

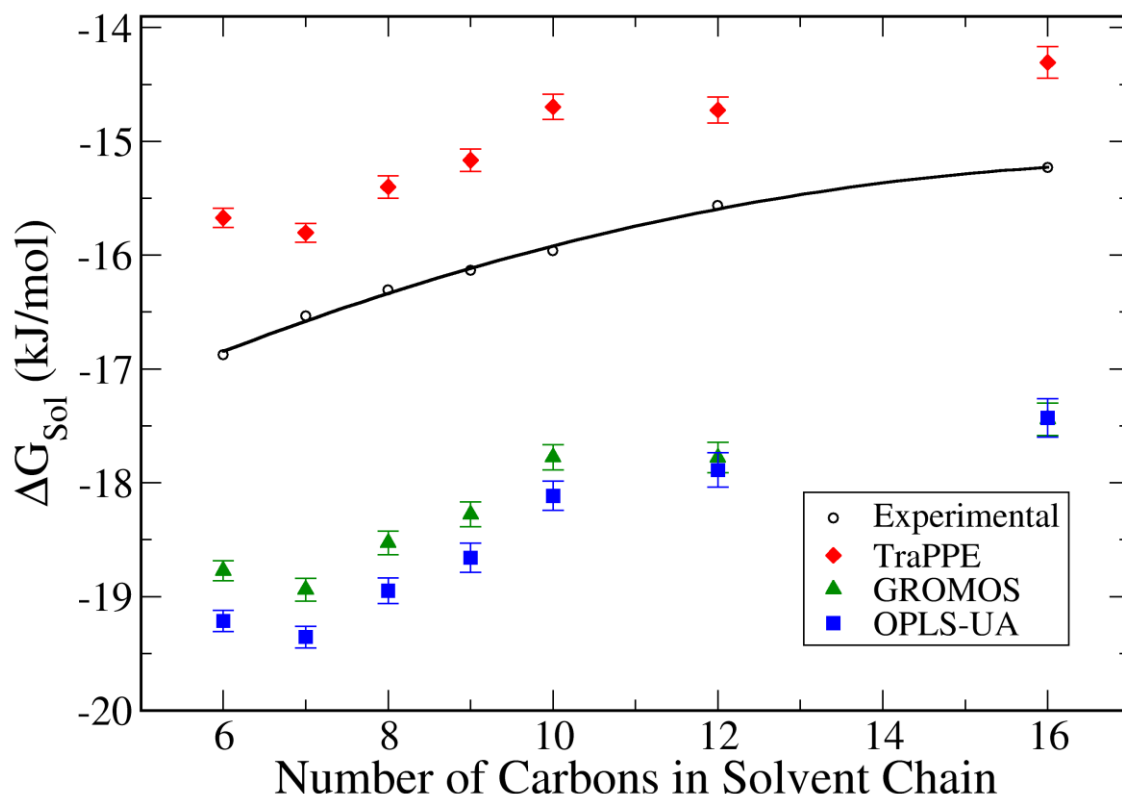


Figure 5 – Solvation free energy of *n*-hexane in linear alkane solvents of increasing chain length. Color code is the same as in Figure 2.

Although in this paper we are mostly concerned with assessing the ability of popular UA models to predict the solvation free energy of alkanes, it is instructive to examine how well they are able to predict properties of pure liquids. One can hypothesize that poorly predicted solvent properties are likely to be reflected in poorly predicted solvation free energies. In Figure 6, we show results for the predicted density and enthalpy of vaporization of pure linear alkane solvents of increasing chain length (corresponding to the solvents used in Figure 5). The enthalpy of vaporization ( $\Delta H_{\text{vap}}$ ) was calculated from the difference in potential energy between a pure liquid simulation and a simulation of a single molecule in vacuum, plus an  $RT$  term. The performance of the three models in predicting experimental values of  $\Delta H_{\text{vap}}$  [54], shown in Figure 6b, is analogous to their performance in solvation free energies – TraPPE slightly underestimates the enthalpy, GROMOS slightly overestimates it, and OPLS-UA shows overestimation that increases with the chain length of the alkane. This suggests that the discrepancies observed in solvation free energy predictions are, at least partly, enthalpic in origin.

The performance rank of the three models with respect to density predictions is also similar to that observed for the free energies, i.e., TraPPE performs best and OPLS-UA performs worse. Interestingly, however, all three models overestimate the experimental density of the pure solvents [55]. This indicates that there is no unique relationship between



the ability of a model to predict density and solvation free energy. In fact, density can potentially be used as a complementary property for force field fitting purposes, together with solvation free energies. We will further explore this possibility in the next paper of this series [25].

It is useful to compare our results for density in the context of the original parameterization of each model, particularly for OPLS-UA and GROMOS, which both used bulk densities for parameter calibration. In the original OPLS-UA paper [19], the authors reported density predictions for linear alkanes up to hexane. Although the match between simulation and experiment for small molecules was reasonable, a deterioration of the predictions with increasing chain length was already evident. In fact, our results for hexane are consistent with those in the original publication, but our data shows that the inaccuracies are amplified for larger alkanes.

In the case of GROMOS, the original paper reports almost exact agreement between simulation and experiment for linear alkanes up to eicosane (C20). The difference with respect to our results for density is completely explained by the different cutoff schemes employed. As we have shown in Figure S3, the results for bulk density are strongly dependent on cutoff radius when long-range dispersion corrections are not applied. Running a simulation for nonane with GROMOS using the exact set of parameters as in the original paper brings the density into perfect agreement with experiment, as expected. Crucially, however, this is not the case for the solvation free energies – a simulation for the self-solvation of nonane using the original GROMOS simulation parameters leads to a prediction of  $-28.71 \pm 0.14$  kJ/mol, compared to  $-28.41 \pm 0.14$  kJ/mol for our choice of cutoff scheme. This difference is very small compared to the variability between different models (see Figure 4) and, more importantly, is in the direction of worse agreement with experiment. As such, the shortcomings of GROMOS in predicting hydrophobic solvation of alkanes cannot be explained by the different cutoff scheme used in their original parameterization.

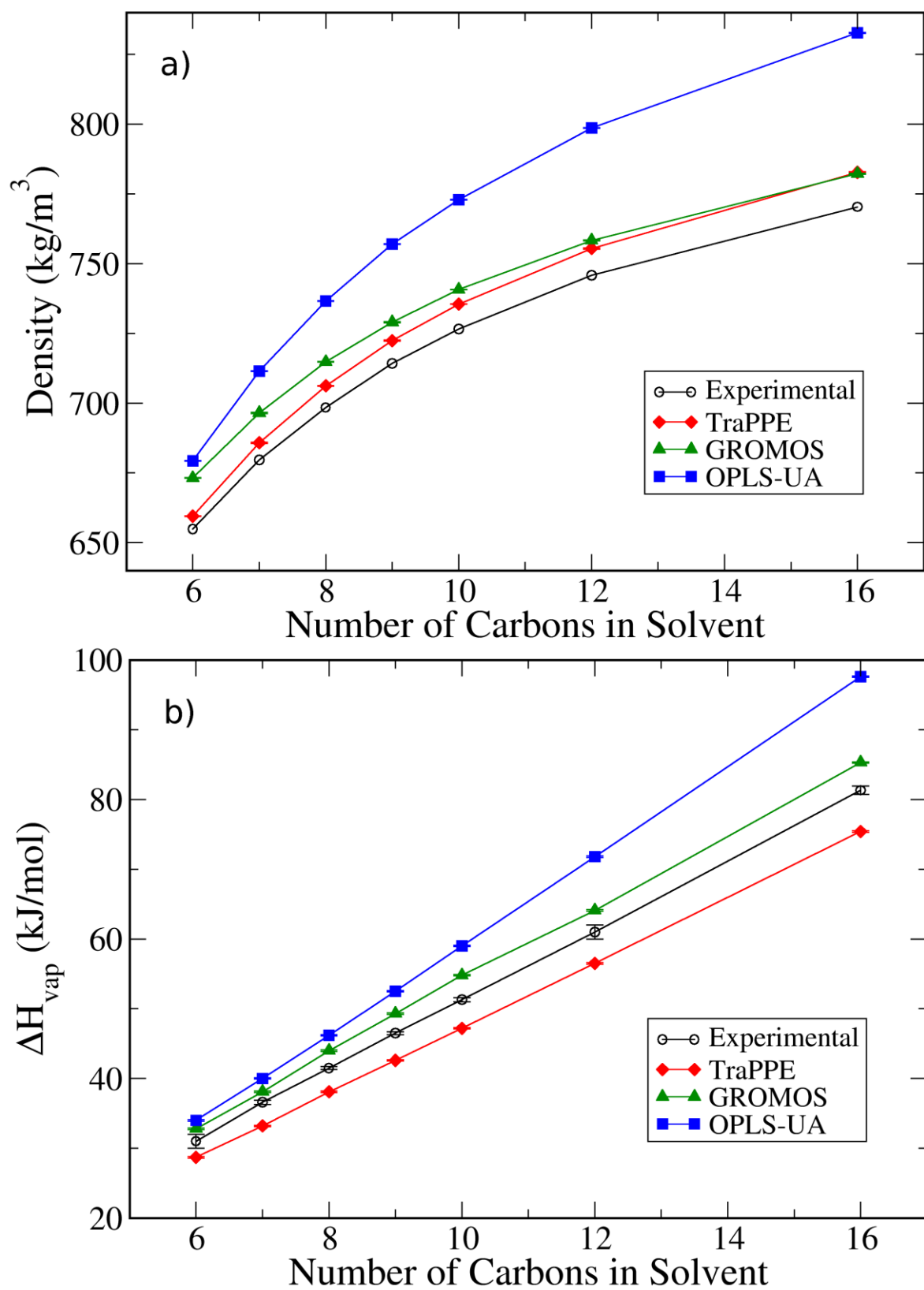


Figure 6 – Density (a) and enthalpy of vaporization (b) of linear alkane solvents of increasing chain length as predicted by different united-atom models, compared to experimental data [54, 55]. Color code is the same as in Figure 2.

### 3.2 – Branched Alkanes

There are fewer experimental measurements for solute/solvent combinations involving branched alkanes [23, 24]. Nevertheless, it is possible to isolate a few trends to assess the performance of the different molecular models. In Figure 7, we plot the solvation free energy of single-branched alkanes (i.e., containing a single CH group) in hexadecane as a function of the number of carbon atoms in the solute (the raw data is shown in Figure S7). The same trends as for linear alkanes were observed – TraPPE underestimates solvation, while GROMOS and OPLS-UA generally overestimate solvation. Interestingly, however, OPLS-UA underpredicts the solvation free energy of isobutane, in an exception to the general trend. This is likely to be due to the parameterization strategy for this model, which considered separate sets of parameters for CH<sub>3</sub> groups in linear, single-branched and double-branched alkanes (see Table 1). An interesting case is that of single-branched hexane isomers, for which we found two data points. The free energy of solvation of 3-methylpentane is slightly higher than that of 2-methylpentane, although the difference is close to the magnitude of the statistical error. Crucially, though, all models are able to capture this trend, suggesting that a UA representation of alkane molecules can describe rather subtle configurational effects.

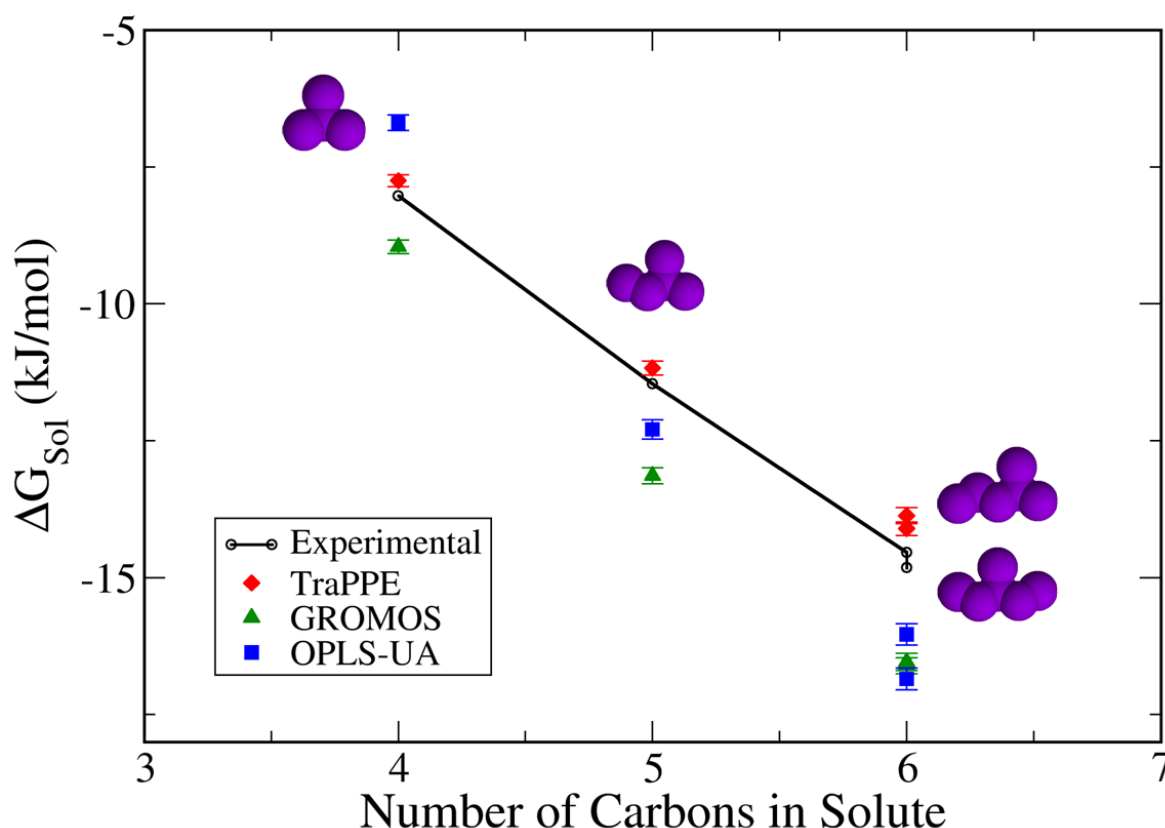


Figure 7 – Solvation free energy of single-branched alkanes of different sizes in hexadecane solvent. For hexane isomers in experiments and all calculations, the point with the lowest free energy corresponds to 3-methylpentane, while the point with the highest free energy corresponds to 2-methylpentane. Color code is the same as in Figure 2.

In Figure S8, we perform a similar test as above, but for double-branched alkane solutes, i.e., containing a single tetrahedrally branched C group (raw data in Figure S9). The trends are essentially the same as those observed in Figure 7, with OPLS-UA once again underestimating the solvation of the smallest solute, neopentane. As argued above, this is likely to be due to the parameterization approach of that force field. Finally, we report in Figure S10 a comparison of predicted free energies for linear alkane solutes in a branched solvent (raw data in Figure S11). In agreement with our observations above, the nature of the solvent has little effect on the trends – TraPPE is shown to underpredict solvation, while GROMOS and OPLS-UA overpredict it.

### 3.3 – *Cyclic Alkanes*

In this section, we compare the three models for their ability to predict solvation free energies involving cyclic alkanes, either as solvents or solutes. When cyclohexane is used as a solvent, the solvation free energy of linear alkanes of increasing chain length (Figure S12, with raw data shown in Figure S13) follows the same trends described above for both linear and branched solvents – the solvation free energy for small solutes is well predicted by all three models, but systematic deviations start to appear beyond C5, with OPLS-UA and GROMOS overestimating solvation and TraPPE underestimating it.

A more insightful test is to consider cyclic alkanes of increasing size dissolved in the same solvent (in this case, hexadecane). The results of this test are shown in Figure 8 (raw data in Figure S14), and reflect the different approaches whereby cyclic molecules are treated in each force field (see section 2.2). All three models do a decent job at predicting the solvation free energy of cyclohexane, which is the cyclic alkane of greatest practical relevance. For larger cycles, the predictions replicate the previously observed trends for linear and branched alkanes (underestimation by TraPPE and overestimation by GROMOS and OPLS-UA). The systematic trend in TraPPE is still present despite the fact that the parameters were optimized for each cycle separately (see Table 1). For cyclopentane, however, this trend is reversed, with TraPPE overestimating solvation. This suggests that the specific parameter values determined for cyclopentane [46] might be somewhat too attractive, contrary to the remaining alkane parameters of this force field. Conversely, the GROMOS model now underestimates solvation of cyclopentane, while OPLS-UA yields a very good prediction of the experimental value, which again indicates that these models can cope well with small alkane molecules.

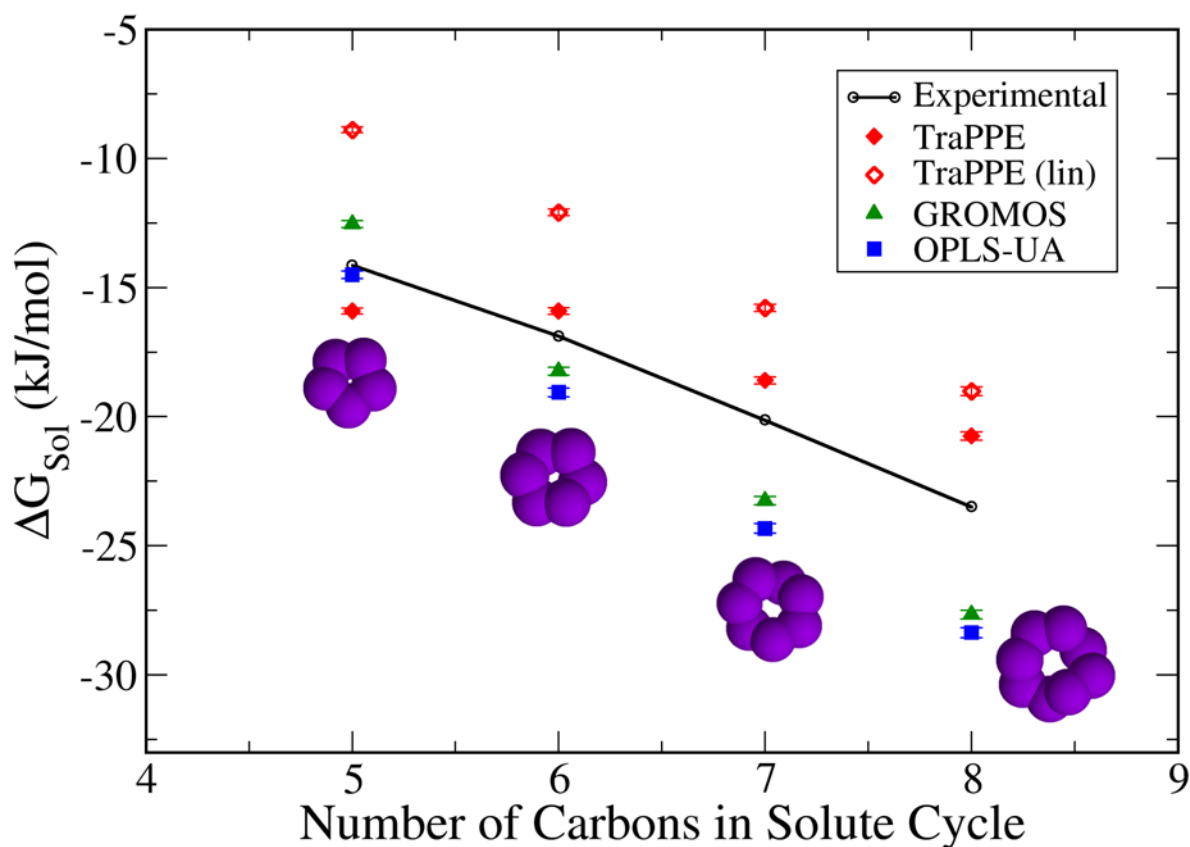


Figure 8 – Solvation free energy of cyclic alkane solutes in *n*-hexadecane solvent. Color code is the same as in Figure 2. Open red diamonds represent calculations with TraPPE assuming the same parameters for linear and cyclic  $\text{CH}_2$  groups.

We have also carried out calculations for the TraPPE model but using the  $\text{CH}_2$  parameters of linear alkanes in cyclic molecules, as was done in the OPLS-UA approach (i.e., assuming that a single set of  $\text{CH}_2$  parameters is applicable to all alkane types). The results of these predictions are shown as open red diamonds in Figure 8. Although the trend with increasing solute size is quantitatively predicted (the slope is very similar to that of the experimental data), the magnitude of solvation is significantly underestimated. This confirms that distinct parameters should be employed for Lennard-Jones centers in cyclic molecules, compared to their linear and branched counterparts, to account for the excluded volume within the ring [44-46]. However, the fact that the trend of free energy with increasing ring size is very accurately described by a single set of cyclic  $\text{CH}_2$  parameters, contrary to what is observed for the recent TraPPE cyclic parameter set, suggests that it may be possible to obtain a single set of cyclic  $\text{CH}_2$  parameters that can describe all cyclic molecules. We explore this issue further in the second paper of this series [25].

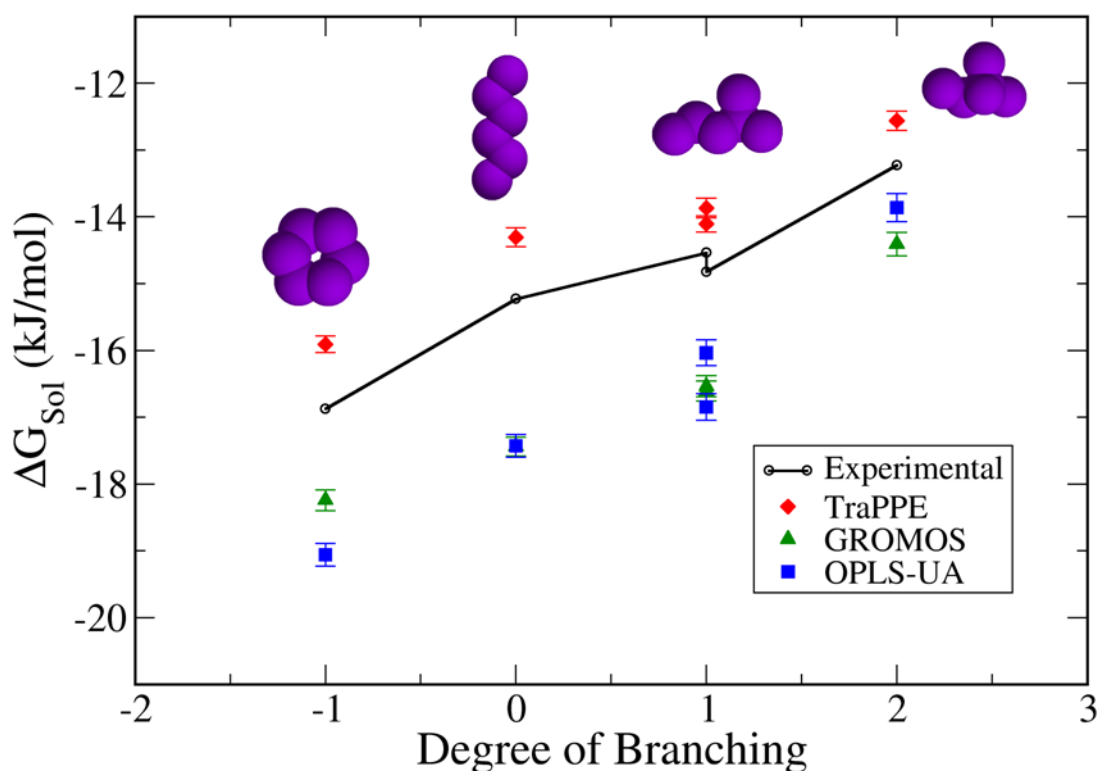


Figure 9 – Solvation free energy of hexane isomers in *n*-hexadecane solvent. Color code is the same as in Figure 2. We have assigned, rather arbitrarily, a degree of branching of -1 to cyclic molecules to allow an easier visualization of the trend.

Thus far, we have mainly discussed the variation of solvation free energy with solute or solvent size. Now that we have presented results for linear, branched and cyclic molecules, it is possible to examine the trends in free energy for alkane isomers. In Figure 9 we report predictions of the three UA models for isomers of hexane dissolved in hexadecane as a function of a somewhat loosely defined “degree of branching” (DOB). We assign a value of 0 to the DOB of linear molecules, 1 to single-branched molecules and 2 to double-branched molecules (this convention would need to be revised when more than one branch is present in a given molecule, but we have avoided such issues here, as the DOB is used mainly for illustration purposes). Furthermore, we rather arbitrarily classify cyclic molecules as possessing a DOB of -1, mainly because this provides a steady increasing trend in the solvation free energy of isomers (see Figure 9). Indeed, the solvation becomes less favorable as the degree of branching increases, with all three models qualitatively capturing this trend (as expected, the magnitude of the free energy is overestimated by GROMOS and OPLS-UA but underestimated by TraPPE).

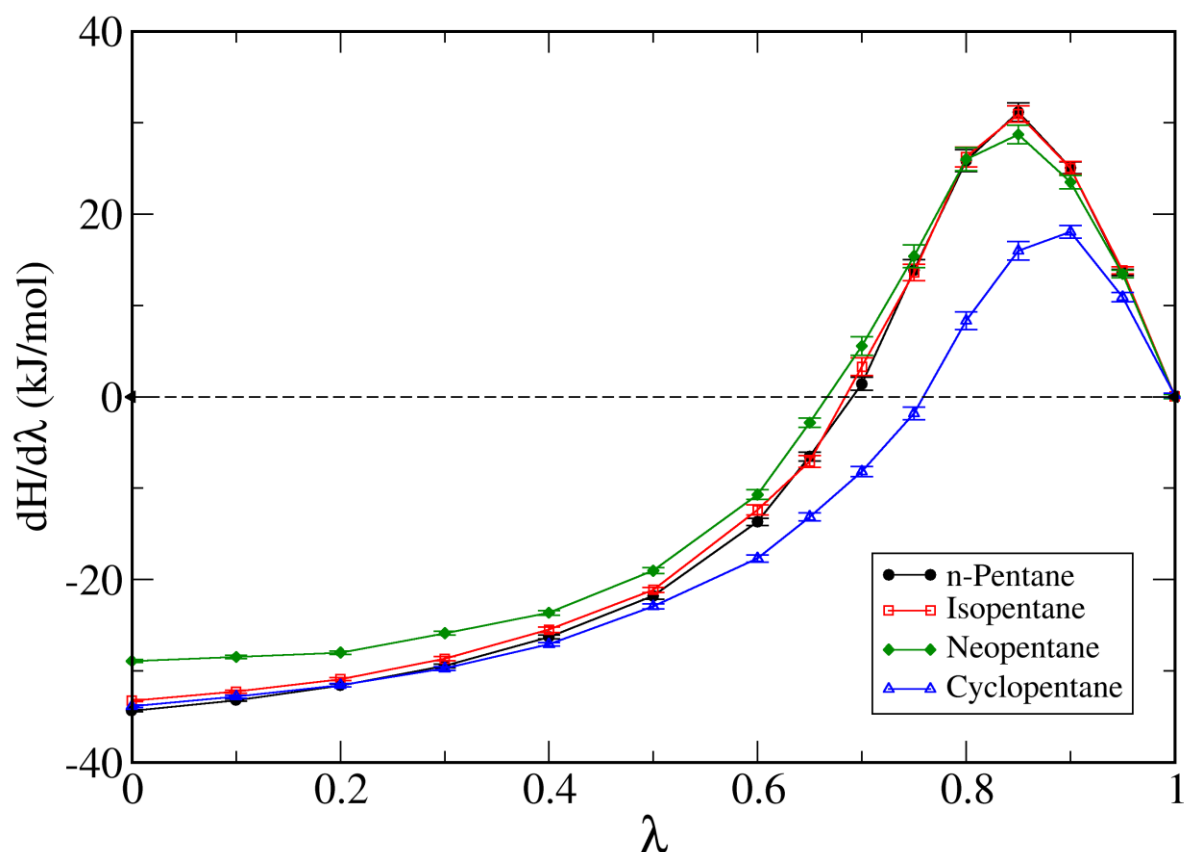


Figure 10 – Gradient of the Hamiltonian with respect to the coupling parameter ( $\lambda$ ) for pentane isomers in *n*-hexadecane solvent, obtained from molecular simulations using the TraPPE model.

More insight into the underlying reasons for the observed trend can be obtained from the Hamiltonian gradient plots for this series of solutes. The raw data obtained with the TraPPE model for hexane isomers is shown in Figure S15, but a clearer picture is seen in Figure 10, which shows the corresponding gradient plots for pentane isomers. The raw data shows that when a linear molecule is transformed into a branched isomer, the dispersive contribution to the solvation free energy increases (low values of  $\lambda$ ). This effect is much more pronounced for the double-branched isomer. Conversely, the cavity formation contribution (high values of  $\lambda$ ) remains nearly the same for the linear and branched isomers. This indicates that despite the more compact form of branched molecules, their excluded volume in the alkane solvent does not change significantly. On the other hand, the presence of additional “buried” Lennard-Jones sites in the form of CH or C groups leads to a loss of attractive interactions with the solvent, as these groups are less exposed to contact with solvent molecules. The net result of these two effects is a decrease in the magnitude of the solvation free energy as the degree of branching of the molecule increases.

Cyclic molecules, on the other hand, show nearly the same dispersion contribution to the free energy as their linear isomers. Indeed, the somewhat smaller  $\sigma$  and larger  $\epsilon$  values for cyclic  $\text{CH}_2$  groups, compared to linear  $\text{CH}_2$  groups (Table 1), compensate almost exactly for

the reduced exposure of cyclic groups to the solvent due to the ring configuration (solvent molecules cannot access the empty space within the ring). However, the cavity formation term is significantly reduced for cyclic isomers relative to their linear and branched counterparts. Although an in-depth investigation of this effect is beyond the scope of this paper, this may suggest that ring-shaped voids have a higher probability of occurrence than rod-shaped (or star-shaped) voids of similar excluded volume in alkane solvents. Alternative explanations based on restructuring of the solvation shells are also possible, as often observed in hydration phenomena due to the hydrogen-bonded nature of liquid water [56], but we believe them to be less likely because of the simplified nature of the non-polar interactions in liquid hydrocarbons.

### 3.4 – Overall Trends

In the final section of this paper, we quantitatively compare the performance of each UA model for predicting solvation free energies of the entire experimental data set considered (results are presented in tabulated form in Supplementary Material, Table S1). In Figure 11, we plot the simulation predictions using OPLS-UA against experimental data for the entire data set, with linear, branched and cyclic solute/solvent pairs denoted by different symbols. The black line represents a perfect match between simulation and experiment, while the red line shows a least squares fit through the data using a linear expression passing through the origin. The slope of this line and the squared correlation coefficient of the fit are shown in the figure legend. Statistical data for the comparison between simulation and experiment are collected in Table 2.

*Table 2 – Measures of deviation between experimental data and simulations using different models, for the entire data set analyzed: MSD = mean signed deviation; RMSD = root mean squared deviation.*

	OPLS-UA	GROMOS	TraPPE
Slope (fit)	1.153	1.131	0.940
R <sup>2</sup> (fit)	0.967	0.986	0.986
MSD (kJ/mol)	2.241	2.111	-0.967
RMSD (kJ/mol)	2.996	2.412	1.204



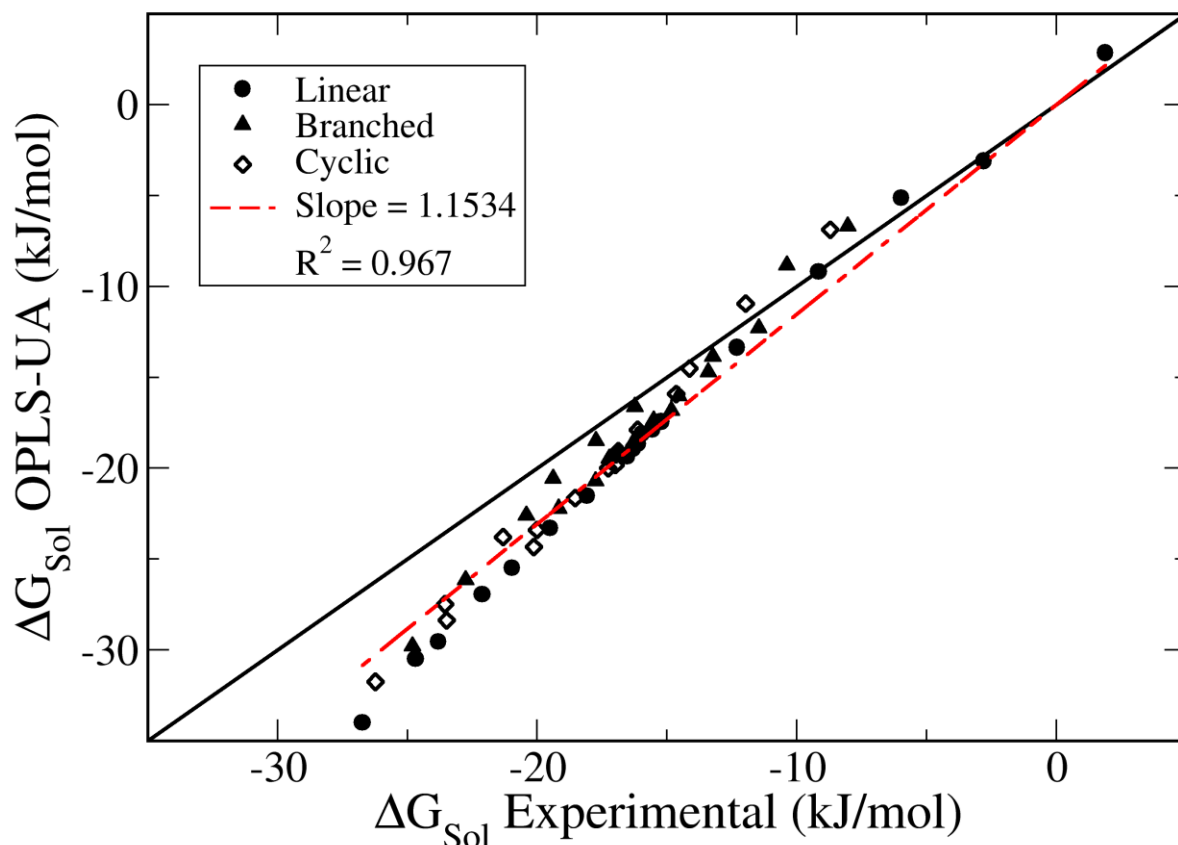


Figure 11 – Comparison between experimental and simulated solvation free energies using the OPLS-UA force field for the entire data set. The dashed red line shows a linear fit (with forced intercept at the origin) through the data. We report also the slope and the correlation coefficient of the fit.

Reflecting our observations above, when the free energy is small in magnitude (corresponding to small alkane solutes), OPLS-UA does a rather good job of predicting solvation. However, for larger alkanes the results deviate significantly and systematically from experiment, leading to a value of the slope above unity. Overall, OPLS-UA overestimates the magnitude of solvation free energies by more than 15%. The correlation coefficient of the fit is relatively poor, reflecting the fact that the data appear to lie on two distinct lines, one for small solutes and another one for large solutes. Compared to the parameters of TraPPE and GROMOS, OPLS-UA has a lower  $\epsilon$  for  $\text{CH}_3$  groups and a higher  $\epsilon$  for  $\text{CH}_2$  groups (the values of  $\epsilon$  for CH and C groups are also higher, but these groups have a relatively small impact on the overall results). In small alkanes, for which OPLS-UA was originally parameterized, these two effects tend to cancel each other out, resulting in predictions that are at least as good as those of the more recent models. However, larger alkanes tend to have a much higher relative proportion of  $\text{CH}_2$  groups compared to terminal  $\text{CH}_3$  groups. In these molecules, the overestimation of the  $\epsilon$  value for  $\text{CH}_2$  groups (and partially also of the CH and C groups) takes hold and leads to a much larger solvation free energy magnitude than in experiment. This highlights the pitfalls of directly transferring force field parameters to different molecules than those originally used in the parameterization. It

also emphasizes the need to consider as broad a range of molecule sizes (and functionalities, although that is not addressed here) when developing force field parameters.

A similar comparison is shown in Figure 12 for GROMOS. Here we observe a systematic overestimation of solvation of about 13% across the entire range of alkanes. Conversely, the results for TraPPE (Figure 13) show a systematic underestimation of solvation by about 6%. Interestingly, the quality of the linear fit for both models was identical, which suggests that the differences are mainly related to the actual values of the force field parameters rather than to any differences in how alkane molecules are described. Comparing the parameters for these two models, we can see that their  $\sigma$  values are quite similar for all types of site. However, the values of  $\epsilon$  in GROMOS are systematically higher (by about 8%) than in TraPPE. This parameter is the most likely candidate to explain the differences between the two models, which suggests that small adjustments of  $\epsilon$  (and perhaps  $\sigma$ ) may be enough to obtain highly accurate solvation free energy predictions. In the second paper of this series [25], we report an attempt at such a refinement.

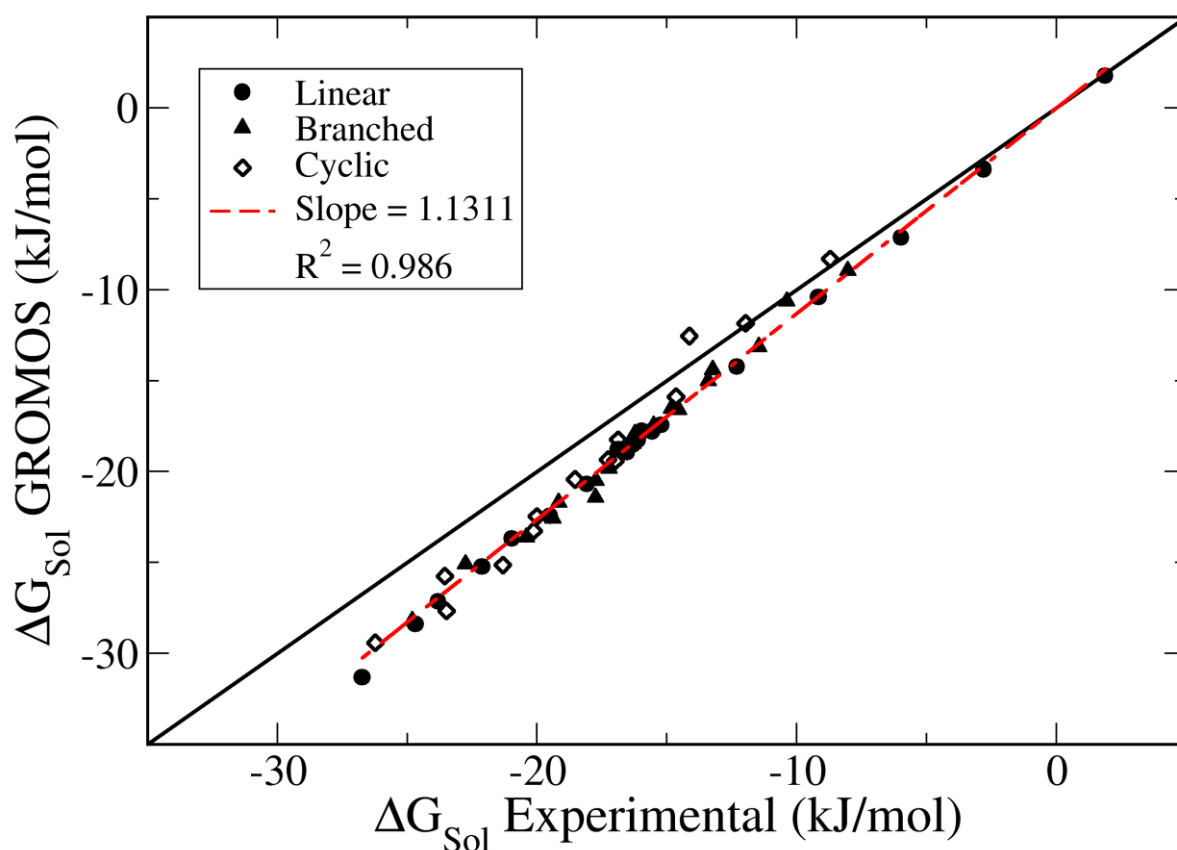


Figure 12 – Comparison between experimental and simulated solvation free energies using the GROMOS force field for the entire data set. Color code is the same as in Figure 11.

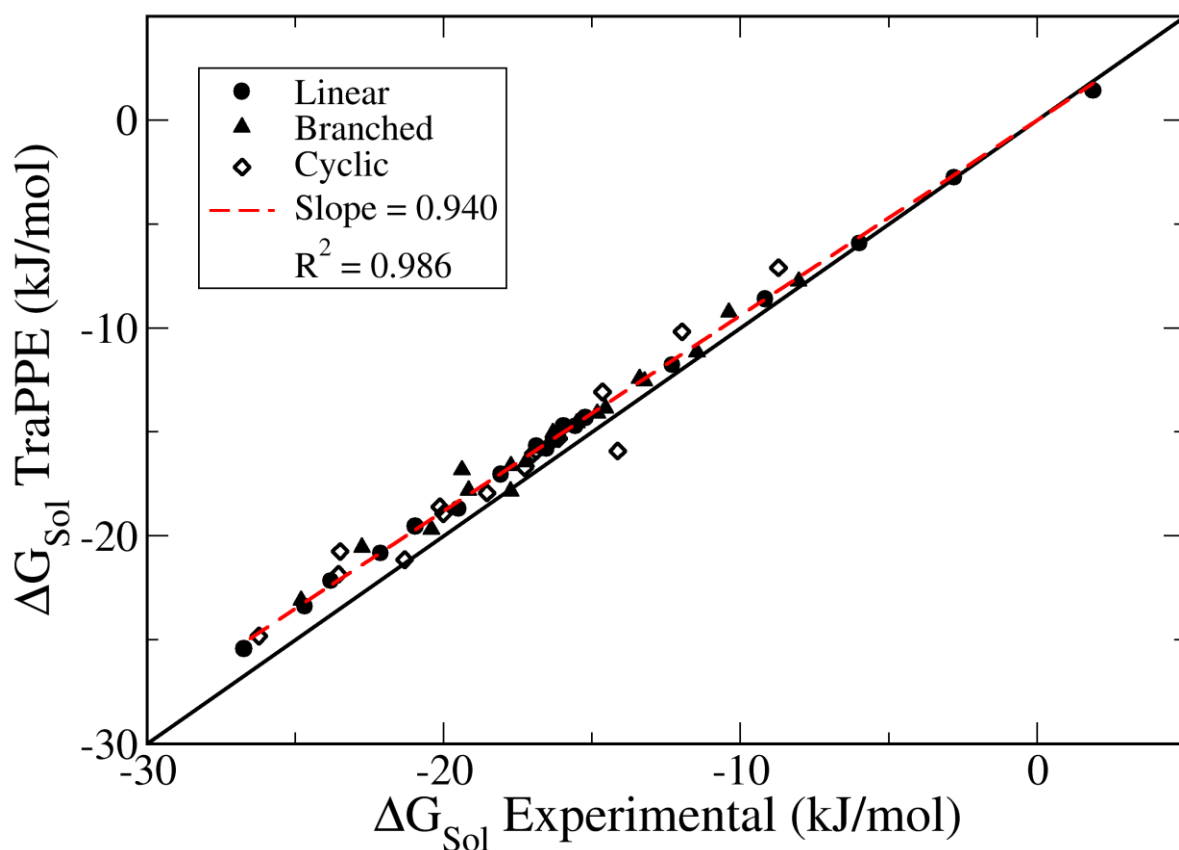


Figure 13 – Comparison between experimental and simulated solvation free energies using the TraPPE force field for the entire data set. Color code is the same as in Figure 11.

#### 4 - Conclusions

We have presented a comprehensive and detailed comparison of three popular united atom force fields (TraPPE, OPLS-UA and GROMOS) for their ability to predict hydrophobic solvation free energies of alkane solutes in alkane solvents. Our results show that all models perform reasonably well for small alkane solutes but systematically fail for larger alkanes. TraPPE underestimates solvation by about 6%, while GROMOS and OPLS-UA systematically overestimate solvation (by 13% and 15% overall, respectively). Average RMSDs between simulation and experiment are all larger than 1 kJ/mol, which is above the estimated statistical uncertainty in experimental data (about 0.8 kJ/mol [57]). We also observed that all three models lead to overestimated densities of pure alkane liquids larger than pentane, although for GROMOS the discrepancy is explained entirely by the use of a different cutoff scheme in the original paper. It is important to emphasize, however, that this difference in cutoff scheme is unable to account for the discrepancies in solvation free energy predictions. The fact that TraPPE underestimates solvation while GROMOS and OPLS-UA overestimate solvation of *all* solutes larger than pentane, leads us to conclude that deviations for all three force fields are systematic in origin – i.e., they are due to the parameters of each force field and not to statistical error in either simulations or experiments. This strongly

suggests that an improved set of UA parameters for alkanes, able to accurately predict hydrophobic solvation free energies, is attainable. In the following paper of this series, we develop and validate a new set of intermolecular parameters for UA alkane molecules.

### Supplementary Material

Additional methodological tests; additional results figures; full table with all experimental and simulated solvation free energies. Input files for all solvation free energy calculations are freely available from the University of Strathclyde's data repository (DOI: 10.15129/4686b937-4e2c-4b02-ae2b-5ac45ade10d6).

### Acknowledgements

The authors would like to thank the volunteer computing platform IBERCIVIS for all the assistance provided during the implementation and execution of project SOLUVEL. Javier Palacios Ramos and Francisco Sanz García deserve a special mention for their tireless support. IBERCIVIS was supported in part by grants from UMIC (*Agência para a Sociedade do Conhecimento*) and FCT (*Fundação para a Ciência e a Tecnologia*) in Portugal, and the IBERCIVIS foundation, CSIC (*Consejo Superior de Investigaciones Científicas*) and *Gobierno de Aragon* in Spain. Special thanks are due to all the volunteers that contributed with their time and computer resources to the IBERCIVIS network. Fruitful discussions with Leo Lue are also acknowledged.

### References

- [1] Ben-Naim, A.; Marcus, Y. Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.* **1984**, *81*, 2016-2027.
- [2] Westergren, J.; Lindfors, L.; Höglund, T.; Lüder, K.; Nordholm, S.; Kjellander, R.; In silico prediction of drug solubility: 1. Free energy of hydration. *J. Phys. Chem. B* **2007**, *111*, 1872-1882.
- [3] Garrido, N. M.; Queimada, A. J.; Jorge, M.; Macedo, E. A.; Economou, I. G. 1-Octanol/Water Partition Coefficients of n-Alkanes from Molecular Simulations of Absolute Solvation Free Energies. *J. Chem. Theory Comput.* **2009** *5*, 2436-2446.
- [4] Rao, S. N.; Singh, U. C.; Bash, P. A.; Kollman, P. A. Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin. *Nature* **1987**, *328*, 551-554.
- [5] Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93*, 2395-2417.
- [6] Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. Alchemical free energy methods for drug discovery: progress and challenges. *Curr. Opin. Struct. Biol.* **2011**, *21*, 150-160.

- [7] Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small molecule hydration free energies in explicit solvent: An extensive test of fixed-charge atomistic simulations. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.
- [8] Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field. *J. Chem. Theory Comput.* **2010**, *6*, 1509–1519.
- [9] Knight, J. L.; Yesselman, J. D.; Brooks, III, C. L. Assessing the quality of absolute hydration free energies among the CHARMM-compatible ligand parametrization schemes. *J. Comput. Chem.* **2013**, *34*, 893–903.
- [10] Zhang, J.; Tuguldur, B.; van der Spoel, D. Force Field Benchmark of Organic Liquids. 2. Gibbs Energy of Solvation. *J. Chem. Inf. Model.* **2015**, *55*, 1192–1201.
- [11] Villa, A.; Mark, A. E. Calculation of the free energy of solvation for neutral analogs of amino acid side chains. *J. Comput. Chem.* **2002**, *23*, 548–553.
- [12] Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656–1676.
- [13] Radzicka, A.; Wolfenden, R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **1988**, *27*, 1664–1670.
- [14] Szklarczyk, O. M.; Bachmann, S. J.; van Gunsteren, W. F. A polarizable empirical force field for molecular dynamics simulation of liquid hydrocarbons. *J. Comput. Chem.* **2014**, *35*, 789–801.
- [15] MacCallum, J. L.; Tieleman, D. P. Calculation of the water-cyclohexane transfer free energies of neutral amino acid side-chain analogs using the OPLS all-atom force field. *J. Comput. Chem.* **2003**, *24*, 1930–1935.
- [16] Chang, J.; Lenhoff, A. M.; Sandler, S. I. Solvation free energy of amino acids and side-chain analogues. *J. Phys. Chem. B* **2007**, *111*, 2098–2106.
- [17] Jorgensen, W. L.; Tirado-Rives, J. The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.
- [18] Garrido, N. M.; Jorge, M.; Queimada, A. J.; Macedo, E. A.; Economou, I. G. Using Molecular Simulation to Predict Solute Solvation and Partition Coefficients in Solvents of Different Polarity. *Phys. Chem. Chem. Phys.* **2011**, *13*, 9155–9164.
- [19] Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. Optimized Intermolecular Potential Functions for Liquid Hydrocarbons. *J. Am. Chem. Soc.* **1984**, *106*, 6638–6646.

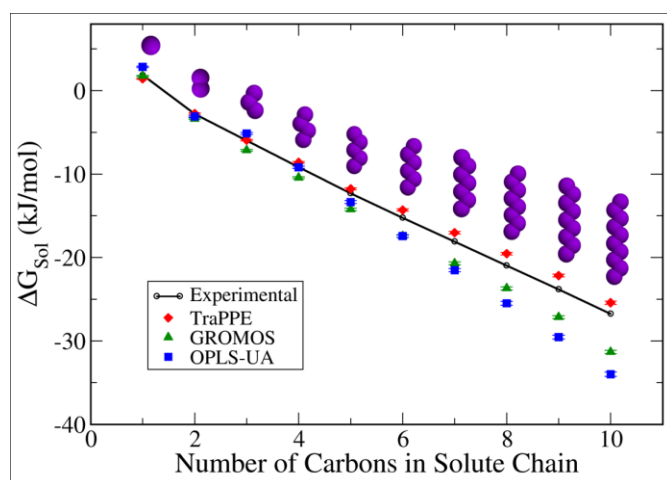
- [20] Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* **2001**, *22*, 1205–1218.
- [21] Martin, M. G.; Siepmann, J. I. Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes. *J. Phys. Chem. B* **1998**, *102*, 2569-2577.
- [22] Martin, M. G.; Siepmann, J. I. Novel configurational-bias Monte Carlo method for branched molecules. Transferable potentials for phase equilibria. 2. United-atom description of branched alkanes. *J. Phys. Chem. B* **1999**, *103*, 4508-4517.
- [23] Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E. Jr. A General Treatment of Solubility. 1. The QSPR Correlation of Solvation Free Energies of Single Solutes in Series of Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794–1805.
- [24] Katritzky, A. R.; Tulp, I.; Fara, D. C.; Lauria, A.; Maran, U.; Acree, W. E. Jr. A General Treatment of Solubility. 3. Principal Component Analysis (PCA) of the Solubilities of Diverse Solutes in Diverse Solvents. *J. Chem. Inf. Model.* **2005**, *45*, 913–923.
- [25] Jorge, M. Predicting Hydrophobic Solvation by Molecular Simulation: 2. New United-atom Model for Alkanes, Alkenes and Alkynes. *submitted*.
- [26] <http://www.IBERCIVIS.com/>
- [27] <http://setiathome.ssl.berkeley.edu>
- [28] <http://folding.stanford.edu>
- [29] Simões, C. J. V.; Rivero, A.; Brito R. M. M. Searching for Anti-Amyloid Drugs with the Help of Citizens: the “AMILOIDE” Project on the IBERCIVIS Platform. *ERCIM News* **2010**, *82*, 25-26.
- [30] Simões, C. J. V.; Rivero, A.; Tarancón, A.; Serrano-Sanz, F.; Romero, J. L.; Abreu, P.; Manuel, C.; Durão, R.; Rosa, J.; Pagaime, J.; Veiga, P.; Jackson, R. M.; Noronha, A.; Vargas, R.; Magalhães, L. T.; Barreira, G.; Brito, R. M. M. Running faster and further together: the AMILOIDE project, the onset of the Portuguese participation on the IBERCIVIS volunteer computing network. *Proceedings of "IBERGRID - 4th Iberian Grid Infrastructure Conference 2010*, Ed. Netbiblo, ISBN 978-84-9745-549-7, pp 404-416.
- [31] Kirkwood, J. G.; Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **1935**, *3*, 300-313.
- [32] Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
- [33] Garrido, N. M.; Jorge, M.; Queimada, A. J.; Economou, I. G.; Macedo, E. A. Molecular Simulation of the Hydration Gibbs Energy of Barbiturates. *Fluid Phase Equilib.*, **2010**, *289*, 148-155.

- [34] Garrido, N. M.; Queimada, A. J.; Jorge, M.; Economou, I. G.; Macedo, E. A. Molecular Simulation of Absolute Hydration Gibbs Energies of Polar Compounds. *Fluid Phase Equilibr.*, **2010**, *296*, 110-115.
- [35] Garrido, N. M.; Jorge, M.; Queimada, A. J.; Gomes, J. R. B.; Economou, I. G.; Macedo, E. A. Predicting hydration Gibbs energies of alkyl-aromatics using molecular simulation: a comparison of current force fields and the development of a new parameter set for accurate solvation data. *Phys. Chem. Chem. Phys.*, **2011**, *13*, 17384-17394.
- [36] Garrido, N. M.; Queimada, A. J.; Jorge, M.; Economou, I. G.; Macedo, E. A. Prediction of the n-hexane/water and 1-octanol/water Partition Coefficients for Environmentally Relevant Compounds using Molecular Simulation. *AIChE J.* **2012**, *58*, 1929-1938.
- [37] Jorge, M.; Garrido, N. M.; Queimada, A. J.; Economou, I. G.; Macedo, E. A. Effect of the Integration Method on the Accuracy and Computational Efficiency of Free Energy Calculations Using Thermodynamic Integration. *J. Chem. Theory Comput.* **2010**, *6*, 1018-1027.
- [38] Shyu, C.; Ytreberg, F. M. Reducing the bias and uncertainty of free energy estimates by using regression to fit thermodynamic data. *J. Comput. Chem.* **2009**, *30*, 2297-2304.
- [39] van Gunsteren, W. F.; Berendsen, H. J. C. Algorithms for Brownian Dynamics. *Mol. Phys.* **1982**, *45*, 637-647.
- [40] Parrinello, M.; Rahman, A. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Phys. Rev. Lett.* **1980**, *45*, 1196-1199.
- [41] van Gunsteren, W.; Berendsen, H. A leap-frog algorithm for stochastic dynamics. *Mol. Simul.* **1988**, *1*, 173-185.
- [42] <http://boinc.berkeley.edu>
- [43] Verlet, L.; Weis, J.-J. Perturbation theory for the thermodynamic properties of simple liquids. *Mol. Phys.* **1972**, *24*, 1013-1024.
- [44] Errington, J. R.; Panagiotopoulos, A. Z. New intermolecular potential models for benzene and cyclohexane. *J. Chem. Phys.* **1999**, *111*, 9731-9738.
- [45] Lee, J.-S.; Wick, C. D.; Stubbs, J. M.; Siepmann, J. I. Simulating the vapour-liquid equilibria of large cyclic alkanes. *Mol. Phys.* **2005**, *103*, 99-104.
- [46] Keasler, S. J.; Charan, S. M.; Wick, C. D.; Economou, I. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria—United Atom Description of Five- and Six-Membered Cyclic Alkanes and Ethers. *J. Phys. Chem. B* **2012**, *116*, 11234-11246.
- [47] Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. A consistent empirical potential for water-protein interactions. *Biopolymers* **1984**, *23*, 1513-1518.

- [48] Horn, H. W.; Swope, W. C.; Pitner, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665-9678.
- [49] Abascal, J. L. F.; Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123*, 234505.
- [50] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926-935.
- [51] Paliwal, H.; Shirts, M. R. Using multistate reweighting to rapidly and efficiently explore molecular simulation parameters space for nonbonded interactions. *J. Chem. Theory Comput.* **2013**, *9*, 4700-4717.
- [52] Little, P. J.; James, M. O.; Foureman, G. L.; Weatherby, R. P.; Bend, J. R. 1-14C-n-hexadecane disposition in the spiny lobster, *Panulirus argus* and the American lobster, *Homarus americanus*. *J. Environ. Pathol. Toxicol. Oncol.* **1986**, *6*, 13-27.
- [53] Winget, P.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Prediction of Vapor Pressures from Self-Solvation Free Energies Calculated by the SM5 Series of Universal Solvation Models. *J. Phys. Chem. B* **2000**, *104*, 4726-4734.
- [54] NIST Chemistry webbook, <http://webbook.nist.gov/chemistry/>, accessed 16/10/2016.
- [55] Weast, R. C.; Astle, M. J. *Handbook of Data on Organic Compounds*. CRC Press: Boca Raton (Fla.), USA, **1985**.
- [56] Chandler, D. Interfaces and the driving force of hydrophobic assembly. *Nature* **2005**, *437*, 640-647.
- [57] Li, J.; Zhu, T.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. Extension of the Platform of Applicability of the SM5.42R Universal Solvation Model. *Theor. Chem. Acc.* **1999**, *103*, 9-63.



## Table of Contents Graphic



This paper presents a systematic comparison of the ability of popular united-atom models to predict solvation free energies of alkanes dissolved in other alkanes. This provides a stringent test of the hydrophobic component of the molecular model. It was found that all models lead to systematic deviations from experiment, which increase with the size of the alkane.

## SUPPLEMENTARY MATERIAL

### **Predicting Hydrophobic Solvation by Molecular Simulation: 1. Testing United-atom Alkane Models**

Miguel Jorge<sup>1\*</sup>, Nuno M. Garrido<sup>2</sup>, Carlos J. V. Simões<sup>3,4</sup>, Cândida G. Silva<sup>3,5</sup>, Rui M. M. Brito<sup>3,5</sup>

*<sup>1</sup>Department of Chemical and Process Engineering, University of Strathclyde, 75 Montrose Street, Glasgow G1 1XJ, United Kingdom*

*Email – miguel.jorge@strath.ac.uk*

*<sup>2</sup>LSRE – Laboratory of Separation and Reaction Engineering – Associate Laboratory LSRE/LCM, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal*

*<sup>3</sup>Chemistry Department and Coimbra Chemistry Centre, Faculty of Science and Technology, University of Coimbra, 3004-535 Coimbra, Portugal*

*<sup>4</sup>BSIM<sup>2</sup> – Drug Discovery, Parque Tecnológico de Cantanhede, 3060-197 Cantanhede, Portugal*

*<sup>5</sup>Center for Neuroscience and Cell Biology, University of Coimbra, 3004-504 Coimbra, Portugal*

## S1 – Computational Methods

For all three force fields tested, non-bonded interactions were modeled by the Lennard-Jones (LJ) potential:

$$E_{ij} = 4\varepsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (\text{S1})$$

where  $r_{ij}$  is the distance between two LJ interaction sites. To determine values of  $\sigma_{ij}$  and  $\varepsilon_{ij}$  for interaction between different atom types (i.e., cross interactions), we applied the combination rules appropriate to each force field – i.e., Lorentz-Berthelot rules for TraPPE and geometric combination rules for GROMOS and OPLS-UA. For completeness, we provide all cross-interaction parameters in Tables S1-S6. The LJ potential can also be expressed in terms of constants  $C_{12}$  and  $C_6$ , which can be easily calculated from the tables of  $\sigma$  and  $\varepsilon$  according to the following relations:

$$C_{12} = 4\varepsilon\sigma^{12}; \quad C_6 = 4\varepsilon\sigma^6 \quad (\text{S2})$$

*Table S1 – Full matrix of Lennard-Jones  $\sigma$  parameters for the GROMOS force field (nm).*

Site	CH <sub>4</sub>	CH <sub>3</sub>	CH <sub>2</sub> (l)	CH <sub>2</sub> (c)	CH	C
CH <sub>4</sub>	0.3710	0.3729	0.3886	0.3831	0.4315	0.4963
CH <sub>3</sub>	0.3729	0.3748	0.3906	0.3850	0.4337	0.4989
CH <sub>2</sub> (linear)	0.3886	0.3906	0.40704	0.4012	0.4520	0.5199
CH <sub>2</sub> (cyclic)	0.3831	0.3850	0.4012	0.3955	0.4455	0.5125
CH	0.4315	0.4337	0.4520	0.4455	0.5019	0.5773
C	0.4963	0.4989	0.5199	0.5125	0.5773	0.6640

*Table S2 – Full matrix of Lennard-Jones  $\varepsilon$  parameters for the GROMOS force field (kJ/mol).*

Site	CH <sub>4</sub>	CH <sub>3</sub>	CH <sub>2</sub> (l)	CH <sub>2</sub> (c)	CH	C
CH <sub>4</sub>	1.2636	1.0468	0.7203	0.7782	0.3463	0.9160
CH <sub>3</sub>	1.0468	0.8672	0.5967	0.6447	0.2869	0.7588
CH <sub>2</sub> (linear)	0.7203	0.5967	0.4105	0.4436	0.1974	0.5221
CH <sub>2</sub> (cyclic)	0.7782	0.6447	0.4436	0.4793	0.2133	0.5641
CH	0.3463	0.2869	0.1974	0.2133	0.0949	0.0070
C	1.2636	1.0468	0.7203	0.7782	0.3463	0.9160

Table S3 – Full matrix of Lennard-Jones  $\sigma$  parameters for the OPLS-UA force field (nm).

Site	CH <sub>4</sub>	CH <sub>3</sub> (l)	CH <sub>3</sub> (e)	CH <sub>3</sub> (sb)	CH <sub>3</sub> (db)	CH <sub>2</sub>	CH	C
CH <sub>4</sub>	0.3730	0.3816	0.3752	0.3819	0.3843	0.3816	0.3790	0.3765
CH <sub>3</sub> (linear)	0.3816	0.3905	0.3839	0.3907	0.3932	0.3905	0.3877	0.3852
CH <sub>3</sub> (ethane)	0.3752	0.3839	0.3775	0.3842	0.3866	0.3839	0.3812	0.3787
CH <sub>3</sub> (single-branch)	0.3819	0.3907	0.3842	0.3910	0.3935	0.3907	0.3880	0.3855
CH <sub>3</sub> (double-branch)	0.3843	0.3932	0.3866	0.3935	0.3960	0.3932	0.3905	0.3879
CH <sub>2</sub>	0.3816	0.3905	0.3839	0.3907	0.3932	0.3905	0.3877	0.3852
CH	0.3790	0.3877	0.3812	0.3880	0.3905	0.3877	0.3850	0.3825
C	0.3765	0.3852	0.3787	0.3855	0.3879	0.3852	0.3825	0.3800

Table S4 – Full matrix of Lennard-Jones  $\epsilon$  parameters for the OPLS-UA force field (kJ/mol).

Site	CH <sub>4</sub>	CH <sub>3</sub> (l)	CH <sub>3</sub> (e)	CH <sub>3</sub> (sb)	CH <sub>3</sub> (db)	CH <sub>2</sub>	CH	C
CH <sub>4</sub>	1.2301	0.9490	1.0322	0.9075	0.8639	0.7793	0.6417	0.5073
CH <sub>3</sub> (linear)	0.9490	0.7322	0.7963	0.7001	0.6665	0.6012	0.4951	0.3914
CH <sub>3</sub> (ethane)	1.0322	0.7963	0.8661	0.7614	0.7249	0.6539	0.5384	0.4257
CH <sub>3</sub> (single-branch)	0.9075	0.7001	0.7614	0.6694	0.6373	0.5749	0.4734	0.3742
CH <sub>3</sub> (double-branch)	0.8639	0.6665	0.7249	0.6373	0.6067	0.5473	0.4506	0.3563
CH <sub>2</sub>	0.7793	0.6012	0.6539	0.5749	0.5473	0.4937	0.4065	0.3214
CH	0.6417	0.4951	0.5384	0.4734	0.4506	0.4065	0.3347	0.2646
C	0.5073	0.3914	0.4257	0.3742	0.3563	0.3214	0.2646	0.2092

Table S5 – Full matrix of Lennard-Jones  $\sigma$  parameters for the TraPPE force field (nm).

Site	CH <sub>4</sub>	CH <sub>3</sub>	CH <sub>2</sub> (l)	CH <sub>2</sub> (c5)	CH <sub>2</sub> (c6)	CH <sub>2</sub> (c7)	CH <sub>2</sub> (lc)	CH	C
CH <sub>4</sub>	0.3730	0.3740	0.3840	0.3805	0.3820	0.3815	0.3810	0.4205	0.5065
CH <sub>3</sub>	0.3740	0.3750	0.3850	0.3815	0.3830	0.3825	0.3820	0.4215	0.5075
CH <sub>2</sub> (linear)	0.3840	0.3850	0.3950	0.3915	0.3930	0.3925	0.3920	0.4315	0.5175
CH <sub>2</sub> (C5 cycle)	0.3805	0.3815	0.3915	0.3880	0.3895	0.3890	0.3885	0.4280	0.5140
CH <sub>2</sub> (C6 cycle)	0.3820	0.3830	0.3930	0.3895	0.3910	0.0070	0.3900	0.4295	0.5155
CH <sub>2</sub> (C7 cycle)	0.3815	0.3825	0.3925	0.3890	0.3905	0.3900	0.3895	0.4290	0.5150
CH <sub>2</sub> (larger cycles)	0.3810	0.3820	0.3920	0.3885	0.3900	0.3895	0.3890	0.4285	0.5145
CH	0.4205	0.4215	0.4315	0.4280	0.4295	0.4290	0.4285	0.4680	0.5540
C	0.5065	0.5075	0.5175	0.5140	0.5155	0.5150	0.5145	0.5540	0.6400

Table S6 – Full matrix of Lennard-Jones  $\varepsilon$  parameters for the TraPPE force field (kJ/mol).

Site	CH <sub>4</sub>	CH <sub>3</sub>	CH <sub>2</sub> (l)	CH <sub>2</sub> (c5)	CH <sub>2</sub> (c6)	CH <sub>2</sub> (c7)	CH <sub>2</sub> (lc)	CH	C
CH <sub>4</sub>	1.2305	1.0013	0.6860	0.7589	0.7329	0.7276	0.7223	0.3198	0.0715
CH <sub>3</sub>	1.0013	0.8148	0.5583	0.6176	0.5964	0.5921	0.5878	0.2603	0.0582
CH <sub>2</sub> (linear)	0.6860	0.5583	0.3825	0.4231	0.4086	0.4057	0.4027	0.1783	0.0399
CH <sub>2</sub> (C5 cycle)	0.7589	0.6176	0.4231	0.4681	0.4520	0.4488	0.4455	0.1973	0.0441
CH <sub>2</sub> (C6 cycle)	0.7329	0.5964	0.4086	0.4520	0.4365	0.0070	0.4302	0.1905	0.0426
CH <sub>2</sub> (C7 cycle)	0.7276	0.5921	0.4057	0.4488	0.4334	0.4303	0.4271	0.1891	0.0423
CH <sub>2</sub> (larger cycles)	0.7223	0.5878	0.4027	0.4455	0.4302	0.4271	0.4240	0.1878	0.0420
CH	0.3198	0.2603	0.1783	0.1973	0.1905	0.1891	0.1878	0.0831	0.0186
C	0.0715	0.0582	0.0399	0.0441	0.0426	0.0423	0.0420	0.0186	0.0042

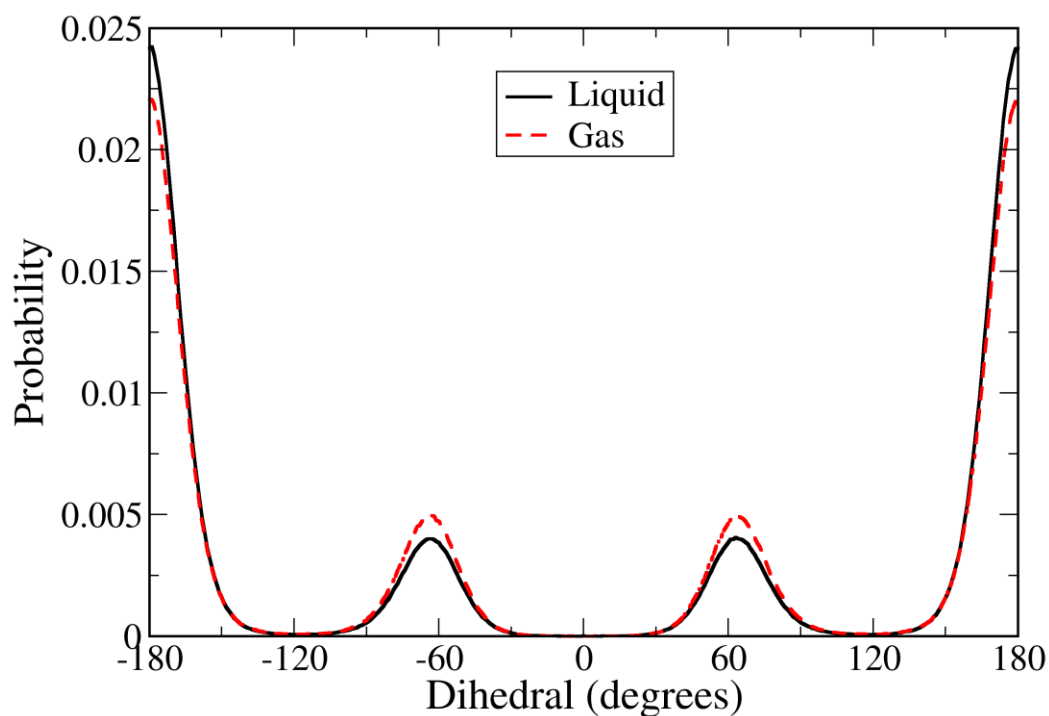


Figure S1 – Dihedral angle distributions for hexadecane in the gas phase (red dashed line) and in the liquid phase (full black line), simulated using the TraPPE model.

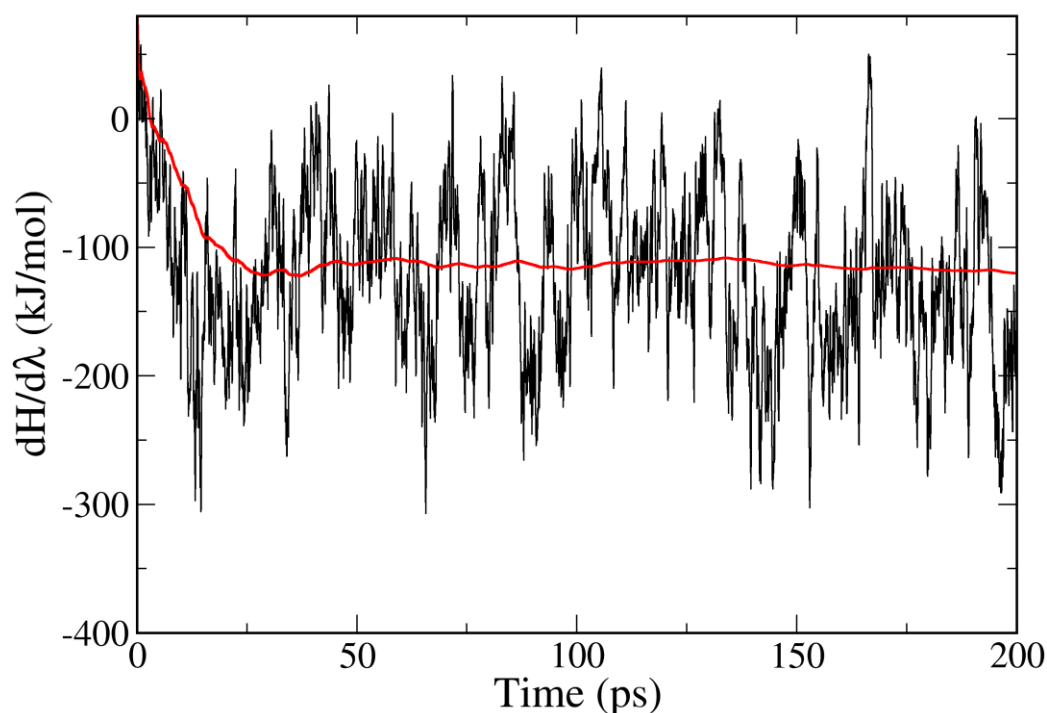


Figure S2 – Convergence of the Hamiltonian gradient (in kJ/mol) over time for an individual configuration. Data is for pentylbenzene in water at  $\lambda = 0.75$ . The full black line shows the raw data points, while the thick red line shows the running average of the data. It is clear that the gradient value converges after about 20 ps.

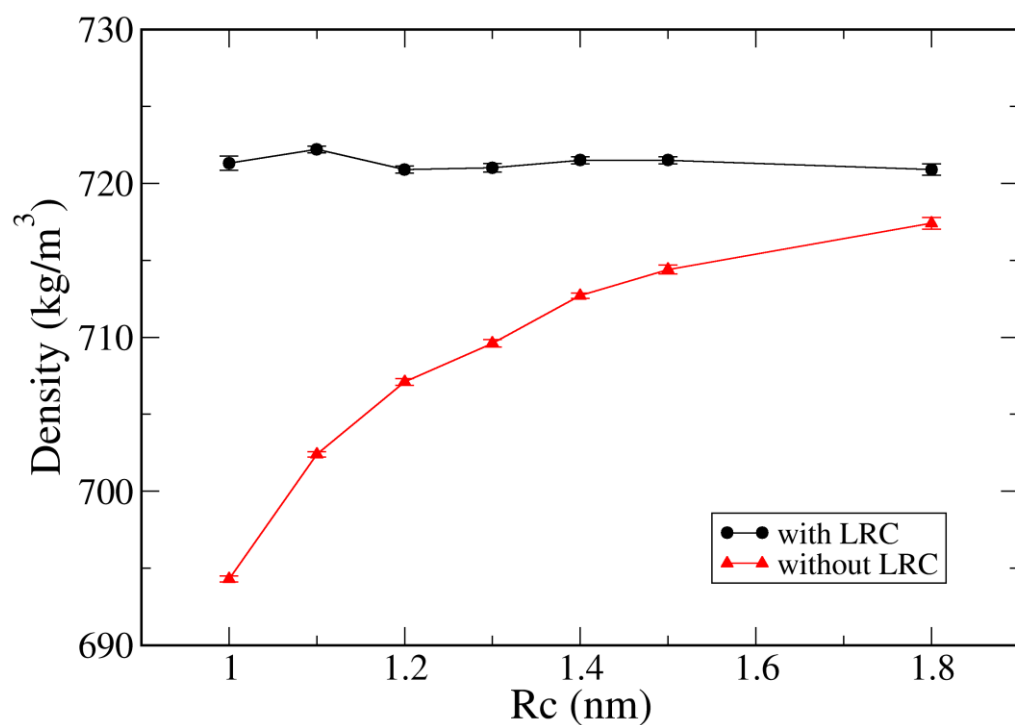


Figure S3 – Effect of cutoff radius on bulk solvent density of *n*-nonane calculated with the TraPPE model. The black circles show results of calculations where long-range dispersion corrections were applied to both energy and pressure, while the red triangles show results obtained without applying long-range corrections.

## S2 – Force-field Comparison

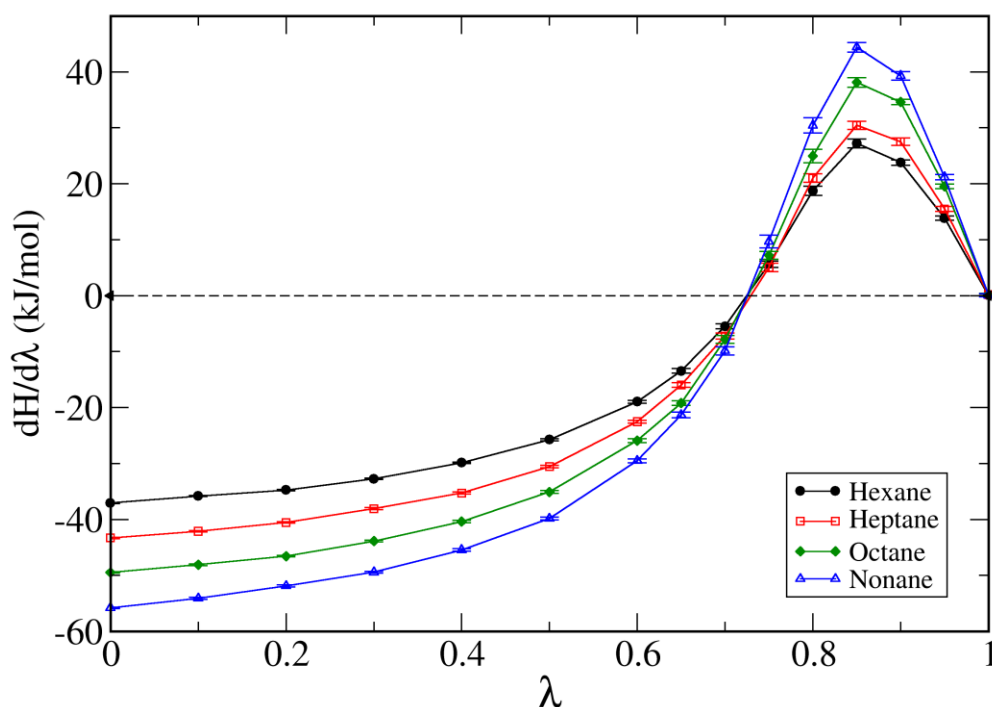


Figure S4 – Gradient of the Hamiltonian with respect to the coupling parameter ( $\lambda$ ) for linear solute/solvent pairs of increasing chain length (i.e., both the solvent and the solute are the same alkane), obtained from molecular simulations using the TraPPE model.

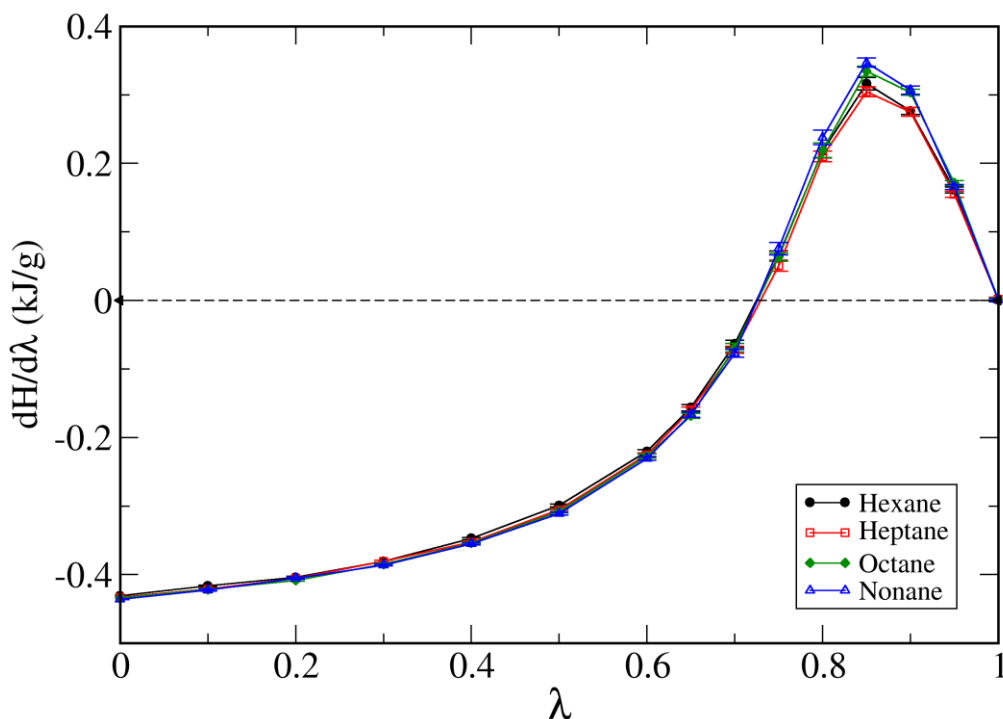


Figure S5 – Gradient of the Hamiltonian with respect to the coupling parameter ( $\lambda$ ) for linear solute/solvent pairs of increasing chain length (i.e., both the solvent and the solute are the same alkane), obtained from molecular simulations using the TraPPE model. Each curve is scaled by the molecular weight of the alkane, such that the units are kJ/g. This scaling makes the data approximately collapse onto a single master-curve.



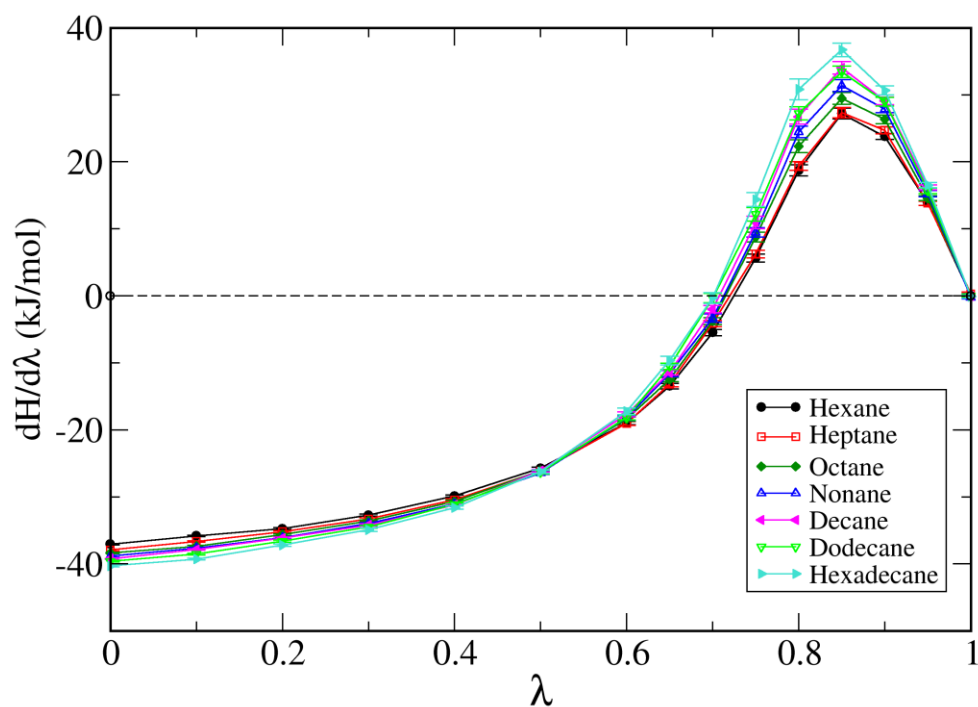


Figure S6 – Gradient of the Hamiltonian with respect to the coupling parameter ( $\lambda$ ) for *n*-hexane in linear solvents of increasing chain length, obtained from molecular simulations using the TraPPE model.

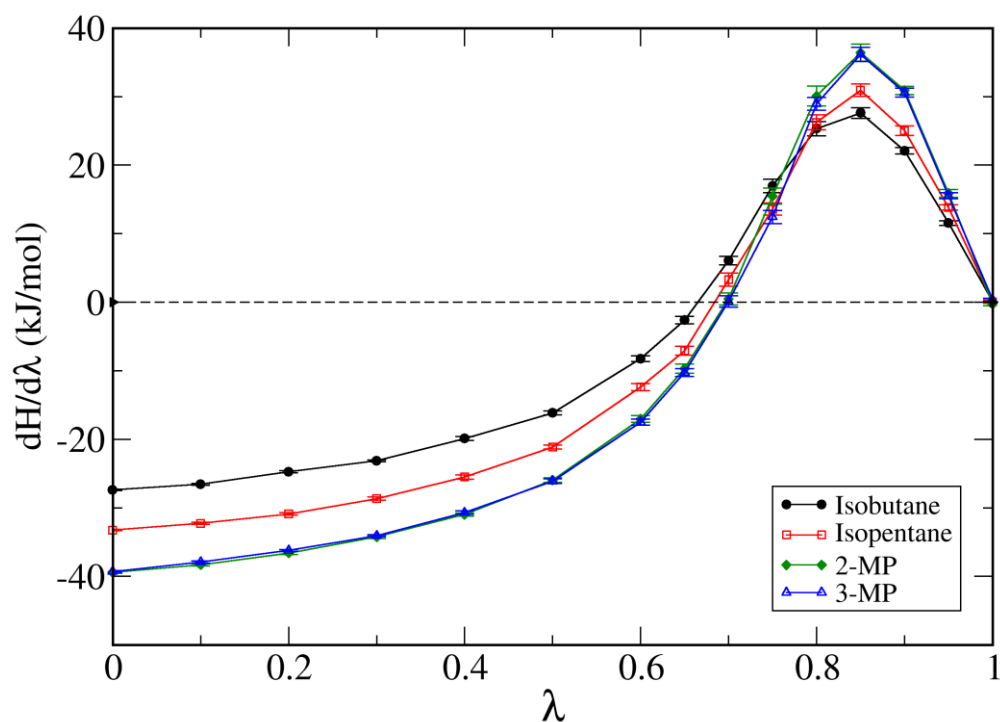


Figure S7 – Gradient of the Hamiltonian with respect to the coupling parameter ( $\lambda$ ) for single-branched alkane solutes in *n*-hexadecane solvent, obtained from molecular simulations using the TraPPE model.

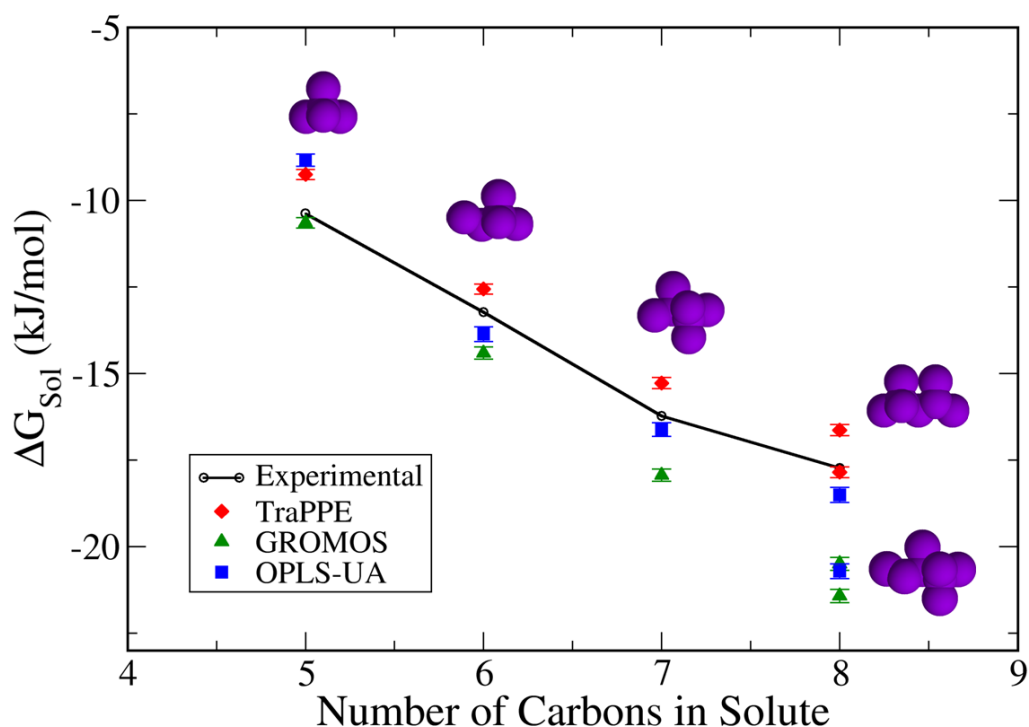


Figure S8 – Solvation free energy of double-branched alkanes of different sizes in hexadecane solvent. For octane isomers in experiments and all calculations, the point with the lowest free energy corresponds to 2,2,3-trimethylpentane, while the point with the highest free energy corresponds to 2,2,4-trimethylpentane.

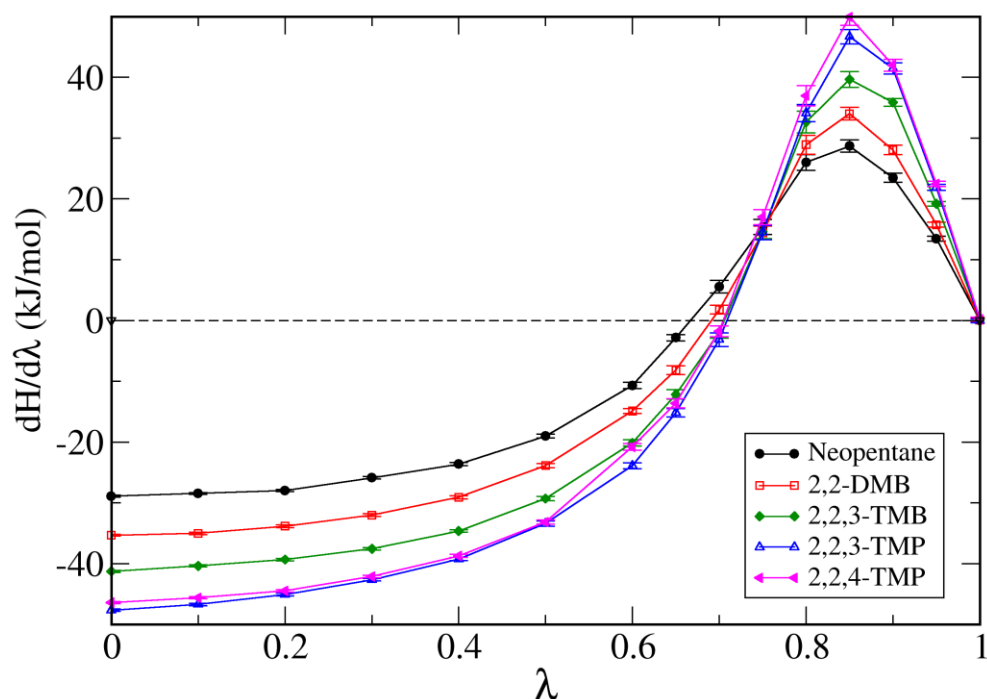


Figure S9 – Gradient of the Hamiltonian with respect to the coupling parameter ( $\lambda$ ) for double-branched alkane solutes in n-hexadecane solvent, obtained from molecular simulations using the TraPPE model.

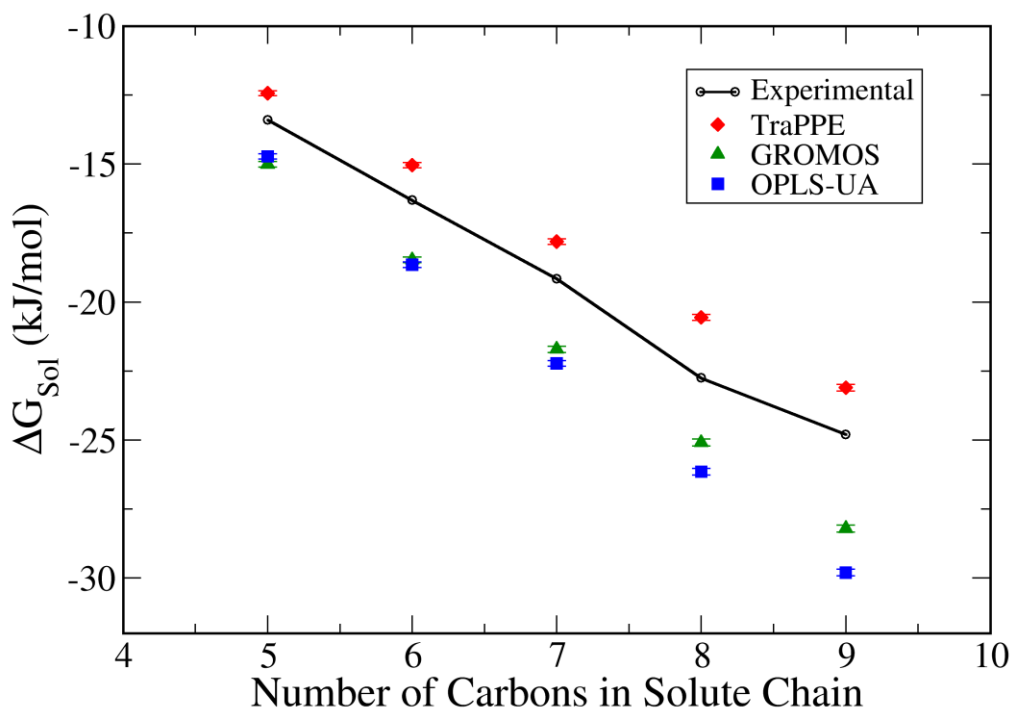


Figure S10 – Solvation free energy of linear alkane solutes in 2,2,4-trimethylpentane solvent. The open circles represent the experimental data [1, 2] while the filled points represent simulations using different models – TraPPE, red diamonds; GROMOS, green triangles; OPLS-UA, blue squares. The black line through the experimental points is a guide to the eye.

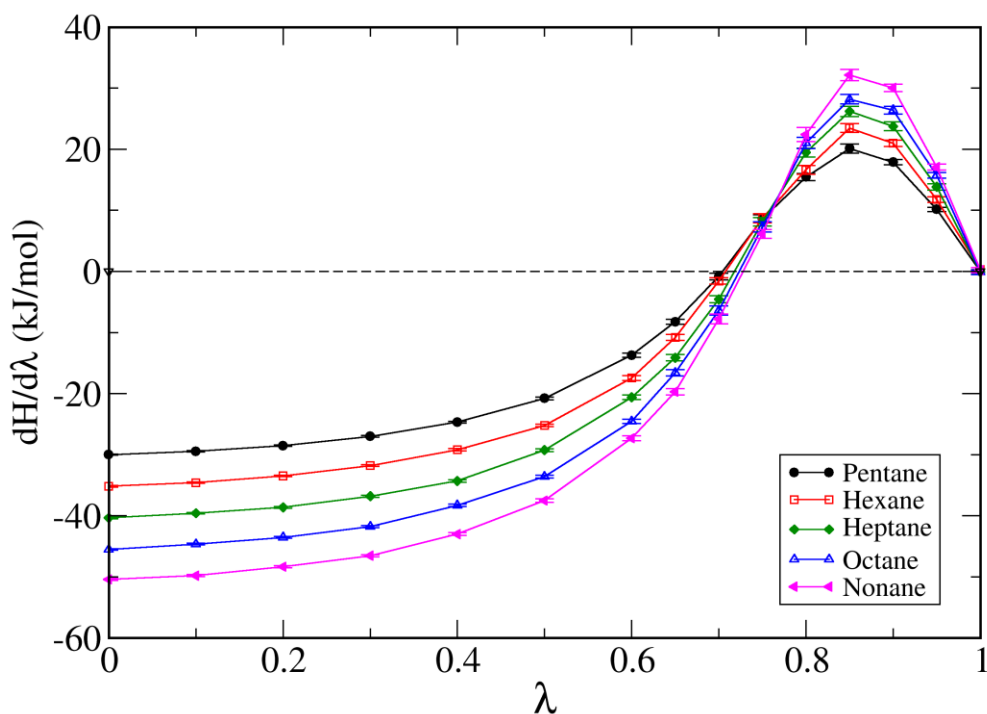


Figure S11 – Gradient of the Hamiltonian with respect to the coupling parameter ( $\lambda$ ) for linear alkane solutes in 2,2,4-trimethylpentane solvent, obtained from molecular simulations using the TraPPE model.

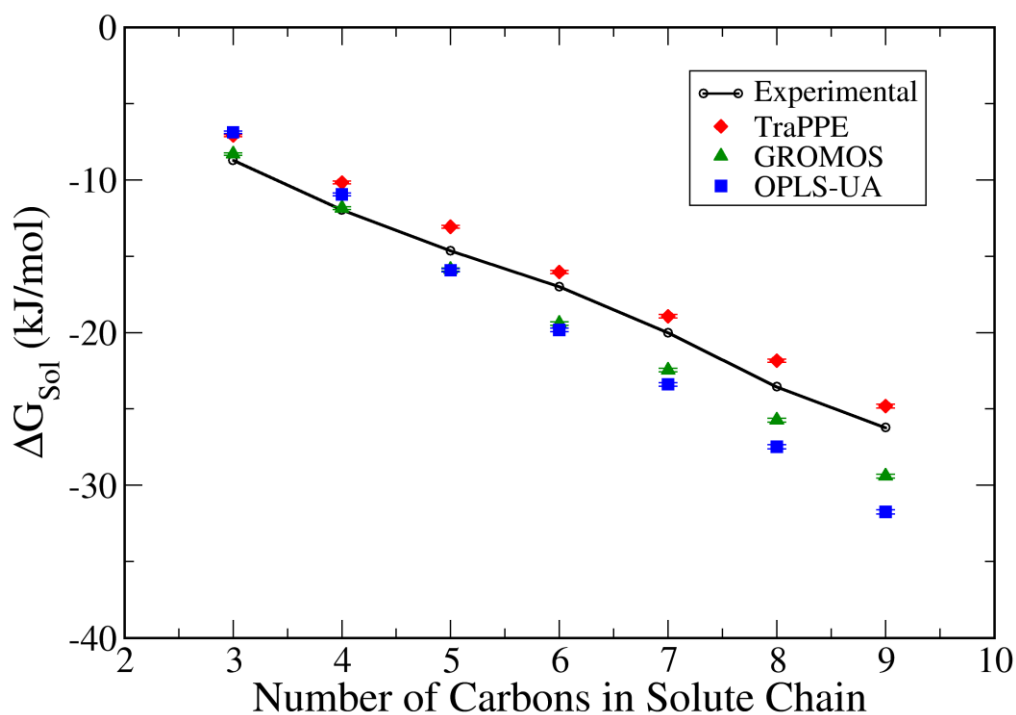


Figure S12 – Solvation free energy of linear alkane solutes in cyclohexane solvent.

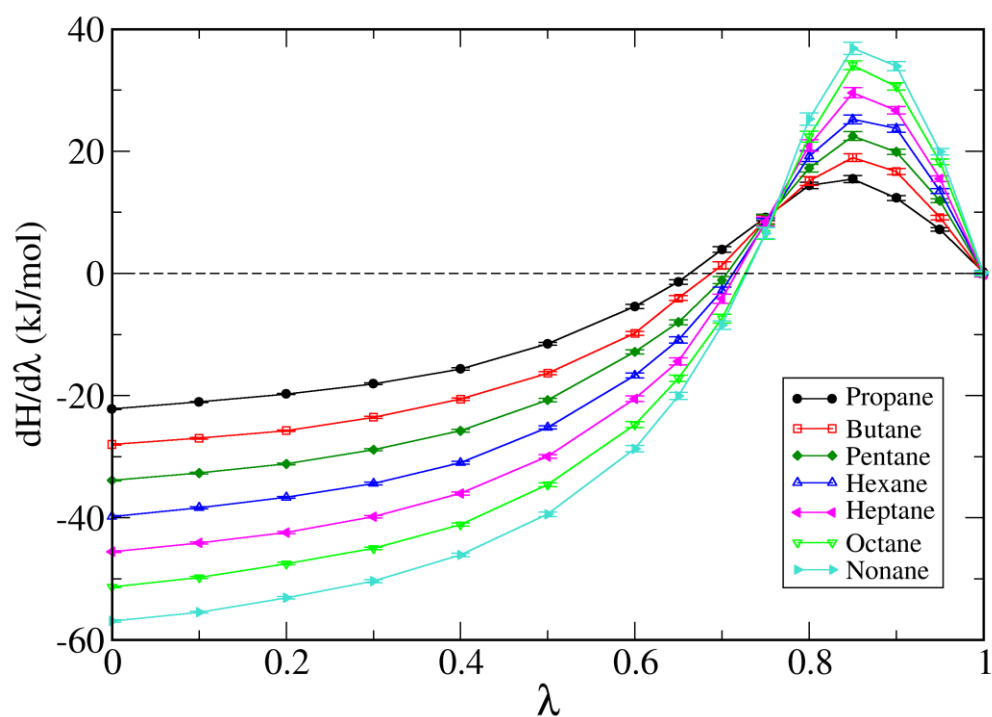


Figure S13 – Gradient of the Hamiltonian with respect to the coupling parameter ( $\lambda$ ) for linear alkane solutes in cyclohexane solvent, obtained from molecular simulations using the TraPPE model.

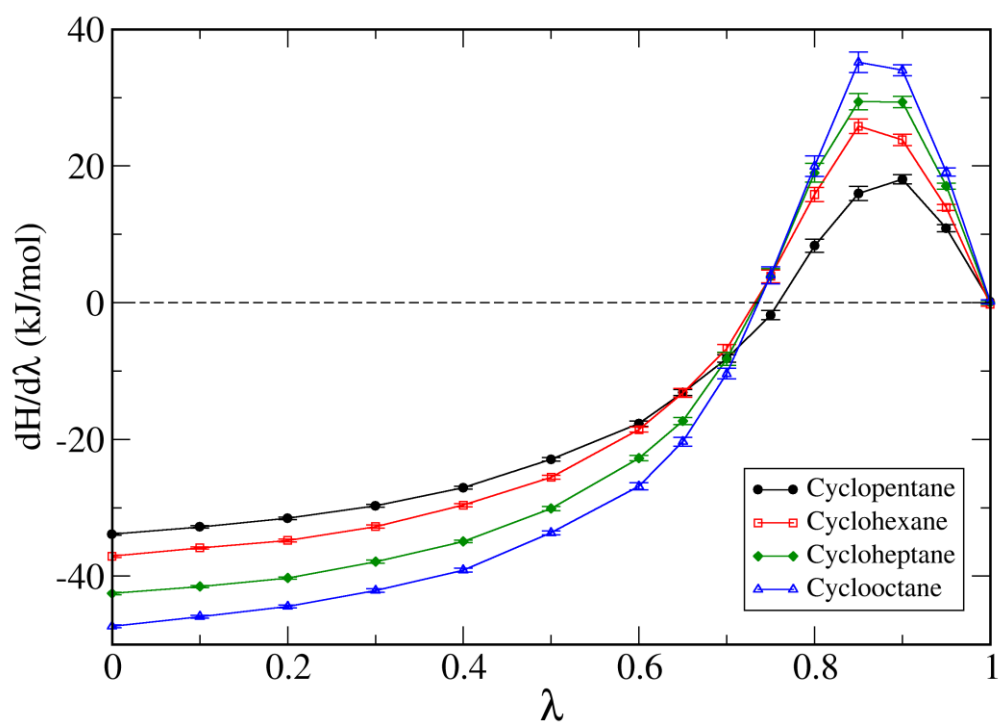


Figure S14 – Gradient of the Hamiltonian with respect to the coupling parameter ( $\lambda$ ) for cyclic alkane solutes in *n*-hexadecane solvent, obtained from molecular simulations using the TraPPE model.

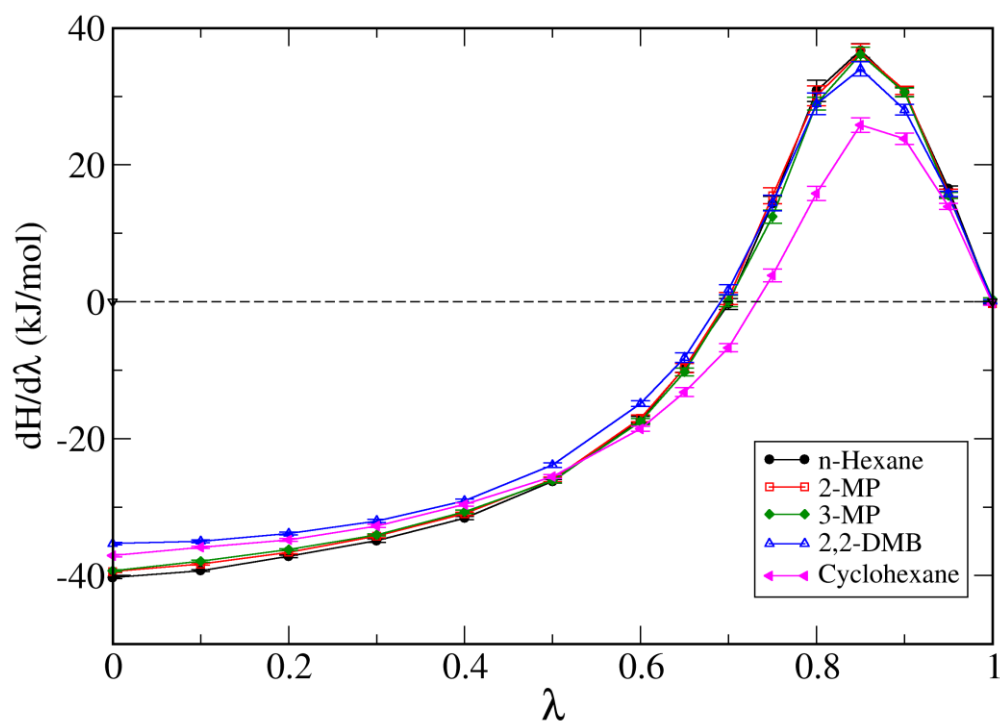


Figure S15 – Gradient of the Hamiltonian with respect to the coupling parameter ( $\lambda$ ) for hexane isomers in *n*-hexadecane solvent, obtained from molecular simulations using the TraPPE model.

Table S7 – Comparison between experimental solvation free energies and those calculated using different models for the entire data set of alkanes examined in this paper. All values are in kJ/mol. Experimental data are from refs [1] and [2]. Uncertainty in the simulated free energies is reported as  $\pm$  the standard error. ASD = absolute signed deviation between simulation and experiment. The first section includes pairs involving only linear alkanes, the second section includes pairs that involve at least one branched alkane, and the third section includes pairs that involve at least one cyclic alkane.

Solute	Solvent	$\Delta G_{\text{exp}}$	OPLS-UA		GROMOS		TraPPE	
			$\Delta G_{\text{sim}}$	ASD	$\Delta G_{\text{sim}}$	ASD	$\Delta G_{\text{sim}}$	ASD
methane	hexadecane	1.88	2.853 $\pm$ 0.074	-0.970	1.760 $\pm$ 0.068	0.123	1.420 $\pm$ 0.061	0.463
ethane	hexadecane	-2.80	-3.095 $\pm$ 0.095	0.292	-3.372 $\pm$ 0.083	0.569	-2.750 $\pm$ 0.100	-0.053
propane	hexadecane	-5.98	-5.132 $\pm$ 0.123	-0.851	-7.149 $\pm$ 0.103	1.166	-5.931 $\pm$ 0.103	-0.052
butane	hexadecane	-9.16	-9.188 $\pm$ 0.142	0.025	-10.421 $\pm$ 0.113	1.258	-8.597 $\pm$ 0.108	-0.566
pentane	hexadecane	-12.30	-13.352 $\pm$ 0.177	1.051	-14.233 $\pm$ 0.152	1.932	-11.771 $\pm$ 0.128	-0.530
hexane	hexadecane	-15.23	-17.430 $\pm$ 0.170	2.200	-17.442 $\pm$ 0.141	2.212	-14.307 $\pm$ 0.140	-0.923
heptane	hexadecane	-18.07	-21.522 $\pm$ 0.193	3.447	-20.693 $\pm$ 0.167	2.618	-17.029 $\pm$ 0.151	-1.046
octane	hexadecane	-20.96	-25.490 $\pm$ 0.227	4.528	-23.696 $\pm$ 0.193	2.734	-19.530 $\pm$ 0.155	-1.432
nonane	hexadecane	-23.81	-29.538 $\pm$ 0.230	5.731	-27.148 $\pm$ 0.187	3.341	-22.160 $\pm$ 0.160	-1.647
decane	hexadecane	-26.74	-33.988 $\pm$ 0.258	7.252	-31.320 $\pm$ 0.217	4.584	-25.418 $\pm$ 0.183	-1.318
hexane	hexane	-16.88	-19.214 $\pm$ 0.094	2.338	-18.772 $\pm$ 0.088	1.896	-15.672 $\pm$ 0.086	-1.204
hexane	heptane	-16.53	-19.355 $\pm$ 0.095	2.821	-18.940 $\pm$ 0.100	2.406	-15.804 $\pm$ 0.084	-0.730
hexane	octane	-16.31	-18.949 $\pm$ 0.113	2.643	-18.529 $\pm$ 0.104	2.223	-15.401 $\pm$ 0.098	-0.905
hexane	nonane	-16.13	-18.659 $\pm$ 0.128	2.524	-18.277 $\pm$ 0.110	2.142	-15.166 $\pm$ 0.098	-0.969
hexane	decane	-15.96	-18.115 $\pm$ 0.128	2.151	-17.776 $\pm$ 0.109	1.812	-14.697 $\pm$ 0.110	-1.267
hexane	dodecane	-15.56	-17.888 $\pm$ 0.151	2.323	-17.778 $\pm$ 0.134	2.213	-14.725 $\pm$ 0.114	-0.840
heptane	heptane	-19.50	-23.298 $\pm$ 0.109	3.800	-22.441 $\pm$ 0.105	2.943	-18.692 $\pm$ 0.095	-0.806
octane	octane	-22.12	-26.933 $\pm$ 0.128	4.812	-25.241 $\pm$ 0.127	3.120	-20.826 $\pm$ 0.109	-1.295
nonane	nonane	-24.69	-30.497 $\pm$ 0.150	5.810	-28.398 $\pm$ 0.137	3.711	-23.388 $\pm$ 0.129	-1.299
isobutane	hexadecane	-8.03	-6.695 $\pm$ 0.142	-1.338	-8.959 $\pm$ 0.122	0.926	-7.753 $\pm$ 0.111	-0.280
isopentane	hexadecane	-11.46	-12.293 $\pm$ 0.174	0.829	-13.141 $\pm$ 0.147	1.677	-11.174 $\pm$ 0.128	-0.290
neopentane	hexadecane	-10.38	-8.840 $\pm$ 0.174	-1.536	-10.648 $\pm$ 0.152	0.272	-9.253 $\pm$ 0.142	-1.123
2-methylpentane	hexadecane	-14.54	-16.036 $\pm$ 0.192	1.498	-16.609 $\pm$ 0.149	2.071	-13.869 $\pm$ 0.147	-0.669

Solute	Solvent	$\Delta G_{\text{exp}}$	OPLS-UA		GROMOS		TraPPE	
			$\Delta G_{\text{sim}}$	ASD	$\Delta G_{\text{sim}}$	ASD	$\Delta G_{\text{sim}}$	ASD
3-methylpentane	hexadecane	-14.82	-16.848±0.198	2.025	-16.538±0.159	1.715	-14.106±0.123	-0.717
2,2-dimethylbutane	hexadecane	-13.23	-13.862±0.212	0.635	-14.411±0.175	1.184	-12.562±0.144	-0.665
2,3-dimethylpentane	hexadecane	-17.22	-19.461±0.199	2.243	-19.825±0.187	2.607	-16.454±0.162	-0.764
2,2,3-trimethylbutane	hexadecane	-16.23	-16.621±0.199	0.387	-17.937±0.178	1.703	-15.278±0.159	-0.956
2,3,4-trimethylpentane	hexadecane	-19.38	-20.567±0.228	1.183	-22.586±0.178	3.202	-16.839±0.176	-2.545
2,2,4-trimethylpentane	hexadecane	-17.73	-18.506±0.215	0.775	-20.498±0.188	2.767	-16.637±0.163	-1.094
2,2,3-trimethylpentane	hexadecane	-17.74	-20.712±0.211	2.972	-21.425±0.192	3.685	-17.851±0.156	0.111
pentane	2,2,4-trimethylpentane	-13.40	-14.723±0.096	1.325	-15.013±0.099	1.615	-12.436±0.086	-0.962
hexane	2,2,4-trimethylpentane	-16.31	-18.645±0.099	2.339	-18.473±0.104	2.167	-15.039±0.093	-1.267
heptane	2,2,4-trimethylpentane	-19.16	-22.222±0.100	3.066	-21.711±0.115	2.555	-17.812±0.101	-1.344
octane	2,2,4-trimethylpentane	-22.75	-26.147±0.119	3.399	-25.091±0.126	2.343	-20.557±0.108	-2.191
nonane	2,2,4-trimethylpentane	-24.80	-29.804±0.120	5.003	-28.205±0.130	3.404	-23.104±0.118	-1.697
2-methylpentane	2,2,4-trimethylpentane	-15.51	-17.366±0.095	1.859	-17.473±0.110	1.966	-14.578±0.093	-0.929
2,3,4-trimethylpentane	2,2,4-trimethylpentane	-20.41	-22.600±0.123	2.189	-23.623±0.128	3.212	-19.695±0.110	-0.716
cyclopentane	hexadecane	-14.14	-14.509±0.145	0.370	-12.546±0.136	-1.593	-15.911±0.111	1.772
cyclohexane	hexadecane	-16.88	-19.061±0.172	2.185	-18.243±0.154	1.367	-15.910±0.124	-0.966
cycloheptane	hexadecane	-20.13	-24.332±0.185	4.206	-23.257±0.162	3.131	-18.594±0.141	-1.532
cyclooctane	hexadecane	-23.49	-28.361±0.196	4.872	-27.664±0.172	4.175	-20.749±0.158	-2.740
cyclohexane	2,2,4-trimethylpentane	-17.27	-20.000±0.099	2.725	-19.336±0.105	2.061	-16.717±0.083	-0.558
propane	cyclohexane	-8.72	-6.883±0.088	-1.840	-8.313±0.077	-0.410	-7.099±0.076	-1.624
butane	cyclohexane	-11.97	-10.948±0.098	-1.025	-11.847±0.088	-0.126	-10.174±0.089	-1.799
pentane	cyclohexane	-14.65	-15.917±0.104	1.265	-15.876±0.099	1.224	-13.075±0.092	-1.577
hexane	cyclohexane	-16.99	-19.825±0.113	2.835	-19.408±0.110	2.418	-16.038±0.103	-0.952
heptane	cyclohexane	-20.01	-23.392±0.118	3.380	-22.462±0.118	2.450	-18.933±0.115	-1.079
octane	cyclohexane	-23.55	-27.484±0.139	3.938	-25.742±0.131	2.196	-21.844±0.115	-1.702
nonane	cyclohexane	-26.23	-31.746±0.146	5.520	-29.408±0.135	3.182	-24.814±0.128	-1.412
2-methylpentane	cyclohexane	-16.13	-17.890±0.119	1.755	-18.051±0.106	1.916	-15.319±0.107	-0.816
2,3,4-trimethylpentane	cyclohexane	-21.32	-23.787±0.126	2.464	-25.136±0.140	3.813	-21.141±0.123	-0.182
cyclohexane	cyclohexane	-18.54	-21.646±0.113	3.111	-20.427±0.118	1.892	-17.931±0.100	-0.604

Table S8 – Comparison between experimental densities [3] and those calculated using different models for the entire data set of alkanes examined in this paper. All values are in kg/m<sup>3</sup>. Uncertainty in the simulated densities is reported as  $\pm$  the standard error.

Solvent	Exp	TraPPE	GROMOS	OPLS-UA
hexane	654.9	659.5 $\pm$ 0.16	673.2 $\pm$ 0.10	679.3 $\pm$ 0.12
heptane	679.7	685.8 $\pm$ 0.23	696.5 $\pm$ 0.19	711.5 $\pm$ 0.14
octane	698.4	706.2 $\pm$ 0.11	714.8 $\pm$ 0.12	736.6 $\pm$ 0.10
nonane	714.2	722.4 $\pm$ 0.25	729.0 $\pm$ 0.26	757.0 $\pm$ 0.12
decane	726.6	735.5 $\pm$ 0.06	740.7 $\pm$ 0.03	772.9 $\pm$ 0.12
dodecane	745.8	755.5 $\pm$ 0.24	758.3 $\pm$ 0.19	798.6 $\pm$ 0.13
hexadecane	770.3	782.7 $\pm$ 0.18	782.2 $\pm$ 0.15	832.7 $\pm$ 0.19
2,2,4-trimethylpentane	687.8	695.8 $\pm$ 0.17	715.6 $\pm$ 0.13	717.2 $\pm$ 0.23
cyclohexane	774.0	791.8 $\pm$ 0.27	771.3 $\pm$ 0.14	809.5 $\pm$ 0.18

## References

- [1] Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E. Jr. A General Treatment of Solubility. 1. The QSPR Correlation of Solvation Free Energies of Single Solutes in Series of Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794–1805.
- [2] Katritzky, A. R.; Tulp, I.; Fara, D. C.; Lauria, A.; Maran, U.; Acree, W. E. Jr. A General Treatment of Solubility. 3. Principal Component Analysis (PCA) of the Solubilities of Diverse Solutes in Diverse Solvents. *J. Chem. Inf. Model.* **2005**, *45*, 913–923.
- [3] Weast, R. C.; Astle, M. J. *Handbook of Data on Organic Compounds*. CRC Press: Boca Raton (Fla.), USA, **1985**.